DUKE
UNIVERSITY

LIBRARY

KE UNIVERSITY RESEARCH STUDIES IN EDUCATION . NUMBER 2

DUKE UNIVERSITY LIBRARY
DURHAM, N. C.

VARIABILITY IN RESULTS FROM NEW-TYPE ACHIEVEMENT TESTS

By EARL V. PULLIAS
Duke University

DUKE UNIVERSITY PRESS
Durham, N. C.
1937

378.756
D847R
no.2

DUKE UNIVERSITY RESEARCH STUDIES IN EDUCATION

VARIABILITY IN RESULTS FROM NEW-TYPE ACHIEVEMENT TESTS

VARIABILITY IN RESULTS FROM NEW-TYPE ACHIEVEMENT TESTS

By EARL V. PULLIAS
Duke University

WITH A FOREWORD BY

WILLIAM A. BROWNELL
Duke University

DUKE UNIVERSITY PRESS
Durham, N. C.
1937

FOREWORD

In this monograph, the second of the Duke University Research Studies in Education, Dr. Pullias presents important new evidence on the variability of so-called objective measures of school achievement. The new-type tests used, both teacher-made and standard, were given under the normal conditions of classroom measurement, and in all probability similar results would be obtained in any school employing comparable tests.

In this monograph the author reports his experimental procedures and his findings at some length—perhaps at greater length than is needed by the reader who is familiar with this type of investigation. Such a reader will be able to find what he wants in the tables and can safely neglect many parts given over to exposition. But Dr. Pullias is writing chiefly for a different audience. He has in mind school officials and teachers who are using new-type tests in ways which may or may not be justified. Such readers require more than a bare outline of what was done and what was found. Accordingly, Dr. Pullias, careful in his experimental work to use only simple techniques which can be duplicated in any school situation, has, in this report, been equally careful to provide enough explanation and interpretation to guarantee real understanding.

Dr. Pullias holds no brief for any particular kind of test. He is concerned only with finding the degree to which different new-type tests, designed to measure the same educational products, agree or disagree in the measures they yield. As a matter of fact, the large amount of disagreement found calls into question a number of educational practices now based upon results from such tests.

In his last chapter Dr. Pullias discusses briefly the factors making for variability in new-type tests. Here the student of educational measurement will find much to challenge his thought. He will welcome too the several testable hypotheses proposed for further research. Some of these hypotheses Dr. Pullias is himself investigating.

William A. Brownell.

ACKNOWLEDGMENTS

In reporting this investigation I wish to express my appreciation to those who have assisted me in the course of its preparation. The criticisms and suggestions of members of the faculty of the Department of Education of Duke University have been very helpful. I am especially indebted to the following persons for assistance in connection with this study: Dr. William A. Brownell, Professor of Educational Psychology, Duke University; Dr. Howard Easley, Assistant Professor of Educational Psychology, Duke University; and Dr. Douglas E. Scates, Director of the Bureau of School Research, Cincinnati Public Schools.

I desire to express my appreciation to Mr. W. F. Warren, Superintendent of Durham City Schools, and to Mr. L. H. Barbour, Superintendent of Durham County Schools, for their permitting me to carry out the study in the city and county schools and for their warm interest in the research. I am grateful to Mrs. J. A. Robinson, Primary Supervisor, Durham City Schools, for her assistance with the administration of the standardized tests, and for her suggestions concerning other aspects of the study. I wish to thank the principals of the schools in which the study was made for their interested co-operation. I am especially indebted to each of the thirty-five teachers, who constructed and administered informal new-type tests, for their efficient and friendly co-operation, without which the teacher-made part of the investigation would have been impossible.

Funds for carrying out some parts of the research here reported were granted by the Duke Research Council. I am deeply grateful for this assistance.

E. V. P.

TABLE OF CONTENTS

PAGE

List of Tables and Figures 9

PART I. INTRODUCTION

CHAPTER

I. The Problem and the Proposed Treatment 13

II. A Brief History of the Problem 19

PART II. THE INVESTIGATION

A. Informal or Teacher-Made Tests

III. Methods Used in the Study of Teacher-Made Objective Tests 26

IV. Correlation Analysis of Results from Paired Informal or Teacher-Made Objective Tests 38

V. Percentile Point Analysis of Results from Paired Informal Objective Tests 48

VI. Teacher-Mark Analysis of Informal Objective Test Results 58

B. Commercial Standardized Tests

VII. Description of Procedures Used in Study of Standardized Tests 65

VIII. Correlation Analysis of Standardized Commercial Test Results 69

IX. Analysis of Standardized Test Findings in Terms of Grade-Equivalent Disparity 74

X. Summary and Conclusions 78

PART III. PRACTICAL IMPLICATIONS AND THEORETICAL PROBLEMS

XI. Educational Implications and Problems for Further Research 81

Appendix 92

Bibliography 98

LIST OF TABLES AND FIGURES

TABLE		PAGE
I.	Percentage Scores on Test V and Test I	29
*II.	Number and Type of Items in Each of Sixty-three Teacher-Made Objective Tests	92
III.	Summary of Facts Which Pertain to Important Characteristics of Teacher-Made Objective Tests	30
IV.	Teachers, Tests, Pupils, and Test Papers Involved in the Study of Teacher-Made Objective Tests Which Cover Shorter Units of Material	34
V.	Teachers, Tests, Pupils, and Test Papers Involved in Study of Teacher-Made Objective Tests Which Cover a Semester of Work	35
VI.	Sample Set of Data Illustrating Types of Analyses Used in the Study of Teacher-Made Objective Tests	36
VII.	Fifth Grade Geography—Correlation between Informal Objective Tests Which Cover Identical Text Material	39
VIII.	Sixth Grade Geography—Correlation between Informal Objective Tests Which Cover Identical Text Material	39
IX.	Fifth Grade History—Correlation between Informal Objective Tests Which Cover Identical Text Material	40
X.	Sixth Grade History—Correlation between Informal Objective Tests Which Cover Identical Text Material	40
XI.	High School History—Correlation between Informal Objective Tests Which Cover Identical Text Material	41
XII.	Summary of Correlations between Tests Which Cover Identical Text Material	42
XIII.	Seventh Grade Geography—Correlation between Informal Objective Tests Which Cover a Semester of Work	43
XIV.	Seventh Grade History—Correlation between Informal Objective Tests Which Cover a Semester of Work	43
XV.	Mean Correlation for All Paired Tests in Geography	45
XVI.	Mean Correlation for All Paired Tests in History	45
XVII.	Summary of Correlations between Paired Tests for All Subjects and All Grade Levels	46
XVIII.	Comparison of Distributions of Correlations for Traditional Tests and for Objective or New-Type Tests	47
XIX.	Fifth Grade Geography—Disparity in Terms of Percentile Points between Informal Objective Tests Which Cover Identical Text Material	49

TABLE		PAGE
XX.	Summary Distribution—Percentile Point Disparity for All Tests Which Cover Identical Text Material	51
XXI.	Summary of Means—Percentile Point Disparity for All Tests Which Cover Identical Text Material	51
XXII.	Summary Distribution—Percentile Point Disparity for All Tests Which Cover a Semester of Work	54
XXIII.	Summary of Means—Percentile Point Disparity for All Tests Which Cover a Semester of Work	55
XXIV.	All Geography Tests—Summary of Mean Percentile Point Disparity	55
XXV.	All History Tests—Summary of Mean Percentile Point Disparity	55
XXVI.	All Tests—Summary Distribution of Percentile Point Disparity	56
XXVII.	All Tests—Summary of Means of Percentile Point Disparity	56
XXVIII.	Fifth Grade Geography—Disparity in Terms of Teachers' Marks between Informal Objective Tests Which Cover Identical Text Material	59
XXIX.	Summary Distribution—Teacher-Mark Disparity for All Tests Which Cover Identical Text Material	60
XXX.	Summary of Means—Teacher-Mark Disparity for All Tests Which Cover Identical Text Material	60
XXXI.	Summary Distribution—Teacher-Mark Disparity for All Tests Which Cover a Semester of Work	61
XXXII.	Summary of Means—Teacher-Mark Disparity for All Tests Which Cover a Semester of Work	62
XXXIII.	All Geography Tests—Summary Distribution of Teacher-Mark Disparity	62
XXXIV.	All History Tests—Summary Distribution of Teacher-Mark Disparity	63
XXXV.	All Tests—Summary Distribution of Teacher-Mark Disparity	63
XXXVI.	All Tests—Summary of Means of Teacher-Mark Disparity	64
XXXVII.	Sub-Test Composition of Metropolitan, New Stanford, and Public School Achievement Tests	66
XXXVIII.	Method of Rotation Used in Administering Standardized Tests	67
XXXIX.	Correlation Table—Reading Tests *a* and *b*	70
XL.	Correlations between Standardized Achievement Tests Which Were Designed to Measure the Same Abilities	71
XLI.	Correlations from Foran's and Loyes' Study of Three Standardized Tests	72

TABLE PAGE

XLII. Mean Correlations for Standardized Tests in Various Subjects Based upon Data from Foran and Loyes and upon Data from the Present Study 72

XLIII. Mean Correlations between American History Tests (Standardized) Based upon Data from Ruch and Others..... 73

XLIV. Arithmetic Computation—Grade-Equivalent Disparity between Tests *a*, *b*, and *c* in Terms of Months..... 75

*XLV. Arithmetic Reasoning—Grade-Equivalent Disparity between Tests *a*, *b*, and *c* in Terms of Months..... 93

*XLVI. Geography—Grade-Equivalent Disparity between Tests *a*, *b*, and *c* in Terms of Months..... 94

*XLVII. Health—Grade-Equivalent Disparity between Tests *b* and *c* in Terms of Months..... 94

*XLVIII. History—Grade-Equivalent Disparity between Tests *a*, *b*, and *c* in Terms of Months..... 95

*XLIX. Language Usage—Grade-Equivalent Disparity between Tests *a*, *b*, and *c* in Terms of Months..... 95

*L. Literature—Grade-Equivalent Disparity between Tests *a* and *c* in Terms of Months..... 96

*LI. Reading—Paragraph-Meaning—Grade-Equivalent Disparity between Tests *a*, *b*, and *c* in Terms of Months..... 96

*LII. Spelling—Grade-Equivalent Disparity between Tests *a*, *b*, and *c* in Terms of Months..... 97

*LIII. Word Meaning—Disparity between Tests *a* and *c* in Terms of Months..... 97

LIV. Summary of Mean Grade-Equivalent Disparity for All Subject Tests 76

FIGURE

I. Percentile Curve for Identical Unit Tests..... 50

II. Percentile Curve for Semester Tests..... 52

*The table appears in the Appendix.

PART I. INTRODUCTION

CHAPTER I

The Problem and the Proposed Treatment

STATEMENT OF THE PROBLEM

This investigation was made to determine the extent to which so-called objective tests produce equivalent results. The method was that of measuring the disparity in the results of comparable objective tests administered under comparable conditions. The findings of the investigation offer a tentative answer to the following question: When two so-called objective tests designed to measure the same thing are used to measure the achievement of the same pupils, to what extent will the results secured by the two measuring instruments vary?

Detailed analysis. The problem may be broken down into more specific queries. 1. If *two competent teachers* construct objective tests designed to measure the acquaintance of a group of pupils with a given section of text material, and if both of these tests are given to the same group of pupils, how will the results of the measurement secured by the application of one test compare with the results when the other test is used?

a. What will be the correlation between the results of the two sets of measurements? When a large number of comparisons are made, what will be the average degree of relationship in terms of a coefficient of correlation?

b. To what extent will the results from the two measurements diverge in terms of percentile points? As an illustration, what per cent of the group will change percentile position as much as twenty percentile points?

c. How much disparity will there be between the results of the two sets of measurements in terms of school marks? For example, how many children will make a mark of *B* on one test and a mark of *A* on the other test?

2. If two or more *standardized* tests designed to measure certain pupil abilities are administered to the same children under comparable conditions how nearly identical will the results of the measurement be?

a. What will be the relationship between the two sets of measures in terms of a coefficient of correlation? What will be the average correlation when a number of subject tests are considered?

b. To what extent will the results of one test vary from the results of another test in terms of grade placement? For example, in the case of how many pupils will the score on one test vary from the score on another test to the extent of making a difference of ten months in grade equivalence? What will be the average amount of disparity in terms of school months for given subject tests?

3. What are the educational implications of the facts revealed in answer to the queries listed under 1 and 2 in the foregoing paragraphs?

4. What are some of the problems for further research which are suggested by the findings of this study?

The problem further analyzed. The objective test received its name from the fact that such tests are relatively objective in regard to scoring. Odell defines such tests in the following manner:

> The term objective is commonly applied to a test which contains exercises of such a sort that there is no disagreement among competent scorers as to what the correct answers are.[1]

That is to say objective tests are free from personal judgment only in respect to scoring. A part of this phase of objectivity is secured because those persons using a test agree to use a given key, as in the case of standardized tests.

But scoring is only one aspect of the total testing situation. In fact, it is a relatively simple aspect of the complex problems involved in the measurement of psychological phenomena. When the subjectivity in scoring is eliminated, do other factors which produce marked variation in test results remain in the testing situation as a whole? If so, what is the extent of the variation? To get a meaningful answer to this question (that is, if one would inquire into the basic elements of objectivity in measurement) one must press his inquiry beyond the relatively superficial matter of uniformity in scoring. The measurement of the results of education involves at least three types of fundamental problem, all of which are closely related to objectivity in the total measurement situation. First, there are the problems relating to the selection of test materials, the formulation of these materials into test items, and the organization of these items into a test. Second, certain important problems grow

[1] C. W. Odell, *Educational Measurement in High School* (New York: The Century Company, 1930), p. 70.

out of the psychological nature of the thing measured, the highly organized, functioning, dynamic mind of the pupil. Third, important issues relating to the interpretation of the results of measurement are involved. Each of these major problems comprises a large number of more specific issues.[2]

This study was not designed as an attack upon these basic causes of variation in social or psychological measurement. On the contrary, the investigation was designed to furnish at least a partial answer to a question which seemed to be a necessary preliminary step to a meaningful attack upon the more fundamental problem; namely, when tests which are objective in regard to scoring are used as the instrument for testing, to what extent do the test results show variation?

INVESTIGATIONS RELATING TO OBJECTIVE TESTS

A complete bibliography of the studies that have been made in an attempt to illuminate various phases of the measurement problem would constitute a volume of considerable proportions. If the field is narrowed to researches relating to the new-type or objective test, the number of titles continues to be large. Fortunately, adequate bibliographies of the research literature pertaining to the objective test are available. The studies pertaining to this particular type of educational measurement which were reported before 1929 were competently reviewed by Ruch.[3] Later researches have been very conveniently summarized by Lee and Symonds.[4] Bibliographies covering the general field of educational measurement are also available.[5]

There have been numerous researches[6] which relate to the reliability of teachers' marks, particularly marks based upon the essay type test, and to the reliability of test scores.

[2] See chap. xi for an extended list of these.

[3] G. M. Ruch, *The Objective or New-Type Examination* (New York: Scott, Foresman and Company, 1929).

[4] J. Murray Lee and P. M. Symonds, "New-Type or Objective Tests: A Summary of Recent Investigations," *The Journal of Educational Psychology,* XXIV (Jan., 1933), 21-38; XXV (March, 1934), 185-191.

[5] C. W. Odell, *A Selected Annotated Bibliography Dealing with Examinations and School Marks* (Bulletin No. 43; Urbana: Bureau of Educational Research, University of Illinois, 1929); H. L. Smith and Wendell William Wright, *Second Revision of the Bibliography of Educational Measurements* (Bulletin of the School of Education, Vol. IV, No. 2; Bloomington: Bureau of Cooperative Research, Indiana University, 1927); "Bibliography on Educational Tests and Their Use," *Review of Educational Research,* III (Feb., 1933), 62-80; Vernon Jones and Robert H. Brown, "Educational Tests," *Psychological Bulletin,* XXXII (July, 1934), 473-499.

[6] See Ruch, *op. cit.,* chaps. iii and xi, for a review of the better known studies.

The typical procedures used in the former type of research—that proposing to determine the reliability of teachers' marks—are, in general, as follows: First, the same teacher scores identical papers at varying intervals;[7] second, a number of teachers mark or evaluate copies of a single pupil's paper;[8] third, a study is made of the distribution of marks given by teachers of different subjects;[9] fourth, marks given to identical children in the same subject during consecutive years are compared.[10] The degree of variation or unreliability is judged by the variation in marks reported. One should note that the first and second procedures involve only the scoring phase of the testing situation;[11] the third method of approach is even less related to the measurement situation as a whole, for it confuses the issue by introducing comparisons between the results of measurement in different subjects; and the fourth type of procedure fails to control properly a number of significant factors, such as important psychological changes which result from a change from one school to another in the case of the pupils acting as subjects. Without doubt many or all of these investigations have contributed something to progress in educational measurement, but they have discovered neither the basic sources of subjectivity in such measurements, nor the degree or extent of variation or disparity when the method of research used was such as to take these fundamental factors into account.

The second type of research on this problem has centered around the more strictly statistical meaning of reliability. Ruch has very aptly described this method of approach:

When two sets of measures of the same ability or function are correlated, we term the resulting coefficient of correlation a reliability coefficient. By "two sets of measures of the same ability or function" we have in mind equivalent or comparable forms of the same test, or some closely analogous pair of measures.

If a teacher gives two forms of a standard test or if she administers two duplicate examinations, the two sets of scores may be compared by

[7] Daniel Starch, *Educational Measurements* (New York: The Macmillan Company, 1918), p. 9.

[8] Daniel Starch and E. C. Elliott, "The Reliability of Grading High School Work in Mathematics," *School Review,* XXI (April, 1913), 254-259.

[9] F. W. Johnson, "A Study of High School Grades," *School Review,* XIX (Jan., 1911), 13-24.

[10] F. J. Kelly, *Teachers' Marks, Their Variability and Standardization* (Teachers College Contributions to Education, No. 66; New York: Teachers College, Columbia University, 1914).

[11] See pp. 14 and 15 for analysis of this term.

correlation, the resulting coefficient in this case being a *reliability coefficient.*

There are several ways of obtaining reliability coefficients when we are studying examinations:

1. Two equivalent (or roughly equivalent) tests may be given and the results correlated.

2. A single test may be given, the papers graded independently by two teachers, and the two sets of marks are then correlated.

3. A single examination, graded by a single person, may be broken into two half-examinations by some chance method (e. g., taking alternate items in each half-form), and the halves are then correlated. . . .

These three methods are not exactly comparable in meaning, but each has its distinct uses in attacking the question of the reliability of examinations.[12]

Some of these studies[13] involved a technique very nearly identical to that used in the portion of this study relating to teacher-made tests, except that these researches were based upon traditional type tests. The coefficients of reliability reported in such studies permit interesting comparisons with the results from the corresponding phase of this study (Chapter IV).

The results from a large number of researches[14] which report reliability coefficients secured by correlating the results from a second administration of the same test, by the administration of comparable forms of a given test, and by correlating two halves of the same test are not directly comparable to the data presented in this study.

ORGANIZATION OF REPORT

The present report has been organized into three major divisions designated as Parts I, II, and III.[15] Part I is introductory in nature. The purpose of this division is to orient the reader in relation to the problem. Part II is a report of the detailed procedures followed, the data secured, the interpretation of these data, and a summary of the conclusions which the data appear to warrant.

[12] Ruch, *op. cit.*, pp. 89-90.

[13] W. S. Monroe and L. B. Souders, *Present Status of Written Examinations and Suggestions for Their Improvement* (Bulletin No. 17; Urbana: Bureau of Educational Research, University of Illinois, 1923).

[14] See Ruch, *op. cit.*, chap. xi, for a review of typical examples.

[15] This report includes with certain modifications research presented to the Graduate School of Arts and Sciences of Duke University as a doctor's dissertation under the title, *Disparity in Results from New-Type or Objective Tests Constructed to Measure the Same Abilities.* The dissertation was done under the direction of Professor William A. Brownell.

Part III is theoretical in nature. The purpose of this division is, in the first place, to present a theoretical discussion of the implications of the findings of this research for educational measurements, and in the second place, to set forth some hypotheses as to the sources of disparity in educational measurements.

Chapter II is presented in order to make the problem here investigated more meaningful. Omission of this chapter would not seriously disturb the continuity of the report of the findings.

CHAPTER II

A Brief History of the Problem[1]

As a result of the close relationship between education and measurement, formal education has been from its inception accompanied by a correspondingly formal type of measurement designed to determine with some degree of immediacy and accuracy the extent to which the purposes of the educative process have been realized. A brief review[2] of some of the prominent examples of this attempt at formal measurement will serve to orient the reader to the research which is the basis of this report.

Review of background factors. Probably all of the highly civilized ancient peoples developed some type of educational measurement. The Chinese boy was required to repeat the classic treatises from memory. The Greek youth was subjected to trying tests in logic, oratory, and related activities. The exact nature of the measures which were used to determine the results of formal, academic training is not always clear. However, the evidence seems to warrant the assumption that varying forms of examinations, oral and written, were used. Ruch comments as follows on this point:

> Oral quizzing, Socratic or otherwise, had been from time immemorial a part of the daily classroom routine; in fact, at times it was all of teaching. Formal written examinations are probably more recent than oral testing, but these date their origins many centuries ago; certainly formal written examinations were firmly intrenched in the educational system of China thirteen hundred years ago, and were familiar to Grecian and Roman teachers.[3]

In America probably the most general method of educational measurement prior to the middle of the nineteenth century was the oral type, for in 1845 Horace Mann[4] made an extended argument

[1] Chapter II, which deals with the historical background of the problem, may be omitted if the reader feels no need for a discussion of background factors.

[2] No extended account of the techniques of measurement or of the application of these procedures to education can be undertaken here. Texts dealing with the history of science should be consulted for the more general development. A serviceable account of the history of the use of measurements in education may be found in C. W. Odell, *Educational Measurement in High School* (New York: The Century Company, 1930), chap. ii.

[3] G. M. Ruch, *op. cit.*, p. 3.

[4] For a valuable detailed analysis of Mann's contribution, *ibid.*, pp. 4 ff.

for the use of written examinations in the public schools of Massachusetts.

An interesting occurrence in the modern testing movement, but one which had relatively little influence, is the work of Reverend George Fisher,[5] an English schoolmaster. Odell makes the following statement concerning the nature and significance of Fisher's work:

> About 1864 he constructed a "scale book" which contained samples of typical questions and of various degrees of proficiency in answering the questions in several school subjects. The questions were intended to be models for the construction of future examinations similar in nature and difficulty. It is not apparent, however, that this work of Fisher's attracted much attention at the time, although it contained the germ of a number of principles later employed.[6]

The work of J. M. Rice seems to be the first direct step in the development of present-day educational measurement. Rice began his researches in 1894 and published his results in a series of articles in the *Forum*.[7] He gave tests in a number of subjects to pupils in different cities, but the portion of this work which is best known is that pertaining to spelling. In this phase of his research Rice gave an identical spelling test to pupils who had been taught spelling in periods of varying lengths and compared the results. This procedure embodied many of the basic elements of later work on standardized tests.

Beginnings of the modern movement. The events just described are of interest from the point of view of history and may have had considerable influence upon the so-called scientific movement in the measurement of educational products. But the fact seems to be, to speak conservatively, that the precipitating cause of the rapid development of this movement was Thorndike's pioneer book[8] published in 1904. This book, due to the author's subsequent influence, definitely committed education to the methods of measurement which had proved so fruitful in the physical sciences. From the publication of this volume the history of measurement in education is an account of the attempts which have been made to realize the ideals set forth therein. In this publication Thorndike presents the sub-

[5] E. B. Chadwick, "Statistics of Educational Results," *The Museum: A Quarterly Magazine of Educational Literature and Science*, III (Jan., 1864), 479-484.

[6] Odell, *op. cit.*, p. 31.

[7] For specific references to these articles, *ibid.*, pp. 31-32.

[8] Edward L. Thorndike, *An Introduction to the Theory of Mental and Social Measurements* (revised ed., 1913; New York: Teachers College, Columbia University, 1904).

stance of his famous creed, which, in certain respects at least, is basic to the subsequent work in the construction of tests and scales. The position expressed in the statement that whatever exists at all exists in some amount and therefore can be measured[9] has been widely accepted in the literature of educational measurements. The following quotations are typical of the manner in which Thorndike further elaborated this creed:

> The scales in actual use in psychology, education, sociology, history and the like are often inadequate in respect to one or more of the essentials of a scale. The work of the student of mental and social measurements is, then, to replace them by better ones so far as he can, to devise methods to make the most out of those which he does not replace, and to avoid attributing to a measurement properties which the scale by which it was obtained does not justify. The last two tasks need no further mention at this point. Concerning the first, it has already been suggested that in cases where quantitative study of human nature and achievement is balked at the very beginning by the lack of series of defined amounts, whose differences from each other and from defined zero points are known, this lack is due rather to lack of study than to any essential insusceptibility of human behavior to rating in units of amount on intelligible scales.[10]
>
> The problem for a quantitative study of the mental sciences is thus to devise means of measuring things, differences, changes and relationships for which standard units of amount are often not at hand; which are variable, and so unexpressible in any case by a single figure; and which are so complex that, to represent any one of them, a long statement in terms of different sorts of quantities is commonly needed. This last difficulty of mental measurements is not, however, one which demands any form of statistical procedure essentially different from that used in science in general.[11]

In reviewing the development of objective tests in education Monroe makes the following statement about Thorndike's early book:

> In 1904 Thorndike published the first edition of his *Mental and Social Measurements.* In addition to an account of statistical procedure, this volume contained many of the principles upon which the construction of our present tests is based. It was revised in 1913, but the revision consists, primarily, of adding concrete illustrations of the principles.

[9] Edward L. Thorndike, "The Nature, Purposes, and General Methods of Measurements of Educational Products," *The Seventeenth Yearbook of the National Society for the Study of Education, Part II* (Bloomington, Ill.: Public School Publishing Company, 1918), p. 16; Walter Scott Monroe, *An Introduction to the Theory of Educational Measurements* (New York: Houghton Mifflin Company, 1923), chap. i; Charles Russell, *Standard Tests* (Boston: Ginn and Company, 1930), chap. iii; and William A. McCall, *How to Measure in Education* (New York: Macmillan Company, 1922), chap. i.

[10] Thorndike, *Mental and Social Measurements,* p. 18.

[11] *Ibid.,* p. 6.

This book is yet an important source of information for workers in this field, and for a number of years was essentially the only source.[12]

The work of Rice shocked professional educators into confusion. His bold exploits into realms hitherto sacred and the disturbing results which he secured threw those educators who were susceptible to the influence of new facts into a state of doubt. If such quantitative results could be obtained in measuring some aspects of education, did it not follow then that all educational products might be evaluated in quantitative terms? The work of Rice constituted a significant practical step, but practical innovations rarely inaugurate movements. Movements have their beginnings in theory—in statements of creed. Such was Thorndike's *Mental and Social Measurements*. This book threw the educational measurements machine into gear, as it were, and it has gathered momentum, in terms of quantity of effort, for the past decade—a momentum the practical results of which have not always been conducive to real progress.[13] Supplied with a theory, workers in education began to produce tests and scales designed to realize the demands of that theory. Odell gives the following description of the early developments:

> It was not until 1908 that anyone followed up Rice's work by publishing a standardized test or scale in any school subject. In this year Stone, a student of Thorndike's, issued his arithmetic reasoning test. This is generally considered the first standardized subject-matter or achievement test. For the next few years standardized tests and scales appeared at the rate of about one a year, practically all of them being by Thorndike and his students. As was probably to be expected, these early tests and scales were for use in subjects entirely or primarily taught in elementary rather than high school. The following list includes those which had appeared by 1913, also one noteworthy but somewhat later one: Courtis Arithmetic Tests, Series A (1909); Thorndike Scale for Handwriting of Children (1909); Hillegas Scale for the Measurement of Quality in English Composition by Young People (1912); Buckingham Spelling Scale (1913); Thorndike Scale for General Merit of Children's Drawings (1913); Ayres Scale for Measuring the Quality of Handwriting of School Children (1912); Ayres Measuring Scale for Ability in Spelling (1915).[14]

These factors (the work of Rice, the theoretical statement of Thorndike, and the production of tests and scales) constituted the more important elements of the scientific movement in education. The general purpose of this movement was to make of education a

[12] Monroe, *op. cit.*, p. 5.
[13] The reference here is to the numerous hurriedly constructed tests and scales and to claims beyond substantiating facts.
[14] Odell, *op. cit.*, pp. 34-35.

quantitative science—that is, a discipline that sought answers to its problems by an appeal to objective fact. Since educationists considered appeal to fact as a method of solving problems to be possible only when measuring instruments capable of revealing these facts are available, much of the energy of the so-called scientific movement in education has been expended in an attempt to produce such instruments. In short, the proposal was to make of education an *objective* science, the chief requisite of such a consummation being the development of *objective* measurements. This term objective when used to characterize a basic and essential factor in the method of approach which would lead to a science of education is identical in meaning with the term as it is used in the physical sciences.[15]

Attack upon traditional type tests. Educationists having agreed upon a method of approach and the goal of a science of education were confronted with the necessity of determining whether or not the methods of measurement which were in use at that time fulfilled the requirements of objective measuring instruments. With this problem as an issue, numerous investigators studied the type of tests then in use, namely, the essay or traditional type. This attack resulted in the production of much evidence[16] in support of the thesis that these instruments of measurement were essentially subjective in their nature and, therefore, unreliable as scientific measuring devices.

Development of objective tests. If the types of measurement in use were inadequate to the needs of a science of education, what then must be done? Obviously, the answer to this query was that more objective measures had to be invented—measures which could be used in the social sciences in a manner comparable to the uses of measurements such as those of length, weight, and time in the physical sciences. When the problem had been thus clearly defined, research workers set themselves earnestly to the task of producing measuring instruments according to the specifications of the older sciences.[17]

[15] Thorndike, *op. cit.*, p. 11.

[16] The specific investigations that produced this evidence have been reviewed too often and too well to demand another presentation here. A good summary may be found in Ruch, *op. cit.*, chap. iii.

[17] The complex and mammoth nature of such an undertaking hardly seems to have occurred, at least consciously, to these workers. It might have been conducive to a more sober development of educational measurements had these facts been taken into account: first, that the social sciences are young; second, that the measuring instruments in use in the physical sciences are the products of a very long and slow development; and third, that probably such slow development could not be profitably short-circuited. For an excellent exposition

Under the pressure of this demand for objective measurement in education, an event occurred which has had a significant influence upon the testing movement. There was developed a type of test item[18] which was called "objective." Apparently intrigued by this term, educationists interested in the measurement of achievement have expended during the past fifteen years a great portion of their energy in the construction, analysis, and praise of these so-called objective tests. The old or traditional-type test became discredited because of its unreliability or subjectivity; the new-type or objective test seemed to fulfill the requirements of science and therefore came to be used very widely and in some cases uncritically. The movement took two directions: first, the development of commercial tests, and second, the use of teacher-made objective tests. The extent and nature[19] of these phases of the commonly called measurement movement in education and the enthusiasm with which they have been prosecuted are matters of common knowledge.

Cause for the popularity of "objective" tests. The cause basic to the rapid production and use of this new-type measurement is significant. A frequent cause for the spread of a method of procedure is the acceptance on the part of those adopting such procedure of assumptions which tend to prejudice them in favor of the method in question. Was there such an assumption in this case? Thorndike and others had proclaimed in very strong terms that objective, scientific measuring devices are basic to a science. A challenge had been made to this youthful discipline, education, that it show the qualities requisite to an entrance into the group of established sciences. One of these qualities, according to Thorndike and others, was an ability to measure objectively phenomena related to its problems in the sense in which the term had been used in relation to the physical sciences.[20] The answer to this demand was the production

---

of this point see Wolfgang Köhler, *Gestalt Psychology* (New York: Horace Liveright, Inc., 1929), chap. ii.

[18] The origin of this type of item, as such, does not seem to be clear. Odell (*op. cit.*, p. 41) makes the following comment on this point: "At the beginning of 1920 appeared an article by McCall, which seems to have been the first published discussion along this line." The general features of this type of test item very likely were patterned after the items used in tests of general intelligence.

[19] See Odell, *op. cit.*, chap. ii; Ruch, *op. cit.*, chap. iv. Almost all of the numerous texts in the field of measurements supply facts bearing on the point.

[20] A. D. Ritchie, *Scientific Method* (London: Kegan Paul, 1923), chap. v; Herbert Dingle, *Science and Human Experience* (New York: Macmillan Company, 1932), chap. vii; Daniel Sommer Robinson, *The Principles of Reasoning, An Introduction to Logic and Scientific Method* (New York: D. Appleton and

and wide use of measuring instruments designated as objective. It seems clear then that an assumption basic to much that has been done and said in relation to measurement in education is that tests which are relatively objective in regard to scoring[21] are therefore objective in the broader sense. A knowledge of the extent of disparity in results from new-type or objective tests should give some indication as to the soundness of this assumption.

Company, 1924), Pt. II; Köhler, *op. cit.*, chap. ii.

[21] This is considered an accurate characterization of the "objective" test. See Charles W. Odell, *A Glossary of Three Hundred Terms Used in Educational Measurement and Research* (Bulletin No. 40; Urbana: Bureau of Educational Research, College of Education, University of Illinois, 1928), p. 43.

PART II. THE INVESTIGATION

(*a.*) INFORMAL OR TEACHER-MADE TESTS

CHAPTER III

METHODS USED IN THE STUDY OF TEACHER-MADE OBJECTIVE TESTS

The problem. The problem of this part of the investigation was to determine the extent to which scores from two objective teacher-made tests tend to be identical when these tests are used to measure the achievement of a group of children with reference to a designated body of subject matter. The problem may be stated as a question: If two competent teachers independently of each other prepare objective tests which they consider accurate measures of pupil knowledge of the same twenty pages in a geography text, and if they give these two tests to the same pupils, how will the results from the two tests compare?

Facts and conditions requisite to a solution of the problem. In order to work toward a solution of the problem, scores from a number of comparable objective tests had to be obtained. It was necessary that competent teachers agree to construct objective tests covering identical or nearly identical subject matter and that these tests be given in pairs to groups of children. The decision was made to secure in a real school situation the data which a solution of the problem demanded. The findings, for this reason, must be interpreted as applying to objective tests when such tests are used by competent teachers in connection with their regular activities in certain subjects.

The school systems. The school systems of the city and county of Durham, North Carolina, were selected for an attack upon the problem. The investigation was made during the school year 1934-35. The school systems in which the study was prosecuted are rated among the most efficient systems of the state. The school populations of the city and county are typical of school populations in general, for there are schools which serve industrial communities, rural sections, and small towns.

No part of the findings of this study should be in any sense considered as a reflection upon the schools in which the data were

gathered. These data should be interpreted as facts which were secured in school systems that are fairly representative of the better public schools of the area in which they are located.

Training and teaching experience of teachers who co-operated. The problem as stated was to compare the results from tests made by *competent* teachers. This is a point of some significance, for if the teachers who constructed the objective tests used in the study were not so well qualified as the average teacher to construct such tests, the validity of any generalization drawn from the data might be questioned.

A summary of the facts pertaining to the academic training and teaching experience of the teachers who constructed the tests used in this study is presented here:

Number of teachers with no degree	9
Number of teachers with Bachelor's degree	26
Number of teachers with Master's degree	6
Number of teachers with Doctor's degree	1
Median years of experience	13

The facts should be noted that twenty-six, or 74.3 per cent, of the thirty-five teachers who constructed the tests had completed sufficient academic study to receive the Bachelor's degree, and that seven, or 20 per cent, had received one or more advanced degrees. No teacher had done less than two years of college work; further, all the teachers who did not have degrees, as well as those who held degrees, had done summer school work at regular intervals during their teaching careers. The median years of experience, thirteen years, seems to indicate that the teachers had ample experience. These facts warrant the conclusion that the teachers who constructed the tests used as a basis for this research were competent public school teachers, if one may judge competency by training and experience.

Attitude of teachers. Although the facts presented in the foregoing paragraph show that the teachers who assisted with this investigation had adequate training and experience, these things alone would not have qualified them to fulfill the purposes of the study. The validity of the findings in such cases depends upon the genuineness of the teachers' efforts.

Those individuals who assisted with this investigation manifested a lively interest in its progress. This interest resulted from several causes, or conditions. First, the teachers felt a need for improving their testing procedures. Second, a personal conference in which

the nature and the purpose of the work were carefully explained was held with each teacher. Third, the marking, scoring, and interpreting of the tests were a part of the regular school work. Fourth, any teacher who for any reason expressed a hesitancy about assisting with the investigation was immediately excused from all obligation with respect to the study. Conscious effort was made to avoid any form of coercion (administrative or otherwise) that might have caused uninterested teachers to co-operate.

GENERAL DESCRIPTION OF PROCEDURE

Example of the technique. The example given here is taken from that part of the investigation carried out in fifth grade history. Teacher V and Teacher I were asked to agree upon a block of text material which would be covered in five weeks of classroom work. The teachers taught this part of the text as they had planned to teach it. Near the end of the teaching period the teachers, independently, constructed objective tests which they considered adequate to measure the pupils' acquaintance with the text material. Both of these tests were mimeographed exactly as the teachers directed; and both tests were given, in this case, to Teacher V's pupils. The two tests were then scored and the scores put into percentage figures in accordance with directions given by the teachers who constructed the respective tests. Thus each pupil had two scores: the one representing his knowledge of the text material as measured by his teacher's (Teacher V's) test, the other representing his knowledge of the same material as measured by Teacher I's test. An example of the scores from the two tests given in the manner described is presented in Table I.

The test constructed by Teacher V is called Test V and that constructed by Teacher I is called Test I. The figures appearing in columns 2 and 3 of Table I are the percentage scores of the group of pupils on the two tests. For example, Pupil 1 made a score of 90 on Test V and a score of 79 on Test I. An examination of the two columns of percentage figures in Table I gives a clear indication of the nature of the original data of this phase of the study.

The tests. Since the investigation was conducted in an actual school situation, there were only two logical criteria for the construction of the tests: (1) that they be the new-type or objective test, and (2) that they be adequate for the desired measurement. Any other requirements as to the nature of the tests would have tended

TABLE I. PERCENTAGE SCORES ON TEST V AND TEST I

Pupil	Score on Test V	Score on Test I
1	90	79
2	85	71
3	70	87
4	92	75
5	95	79
6	67	71
7	92	75
8	60	67
9	37	54
10	92	71
11	80	71
12	92	79
13	72	75
14	97	67
15	72	66
..	..	..
..	..	..
..	..	..
40	85	62
41	75	71
42	60	75

to formalize the study and to divorce it from the actual school situation.

The tests were composed of commonly used types of item. The first items from one randomly selected test in each subject field are presented here in order to give an indication of the types of item involved.

Geography (fifth grade): When water changes to vapor we call the process (condensation, evaporation, moisture).

Geography (fifth grade): In coming from the Congo to the United States you would cross the ..... Ocean.

Geography (seventh grade): We are what we are largely because of where we are. (True-False)

History (fifth grade): The story of our nation begins more than (100, 300, 400, 800) years ago.

History (seventh grade): The Greeks had a strong central government. (True-False)

History (high school): .... Tariff of Abomination, ... Compromise Tariff of 1833, ... Force Act, ... South Carolina Nullification Ordinance. (To be arranged in chronological sequence).

Health (sixth grade): Proteins are found in such foods as ....., ..., ...

The sample items presented show only a part of the types used, but this chance sample will give a relatively clear indication of the nature of the tests.

The average length of the tests used in this study was approximately forty items. Some of the tests were relatively short, but two facts should be kept in mind in this connection: first, the tests were considered adequate by competent teachers, and second, these shorter tests covered brief units of material.[1]

More detailed analysis of the tests. The purpose of this section is to provide detailed information concerning the nature of the tests used in this part of the study. These facts appear in Table II.*[2] The most important facts in Table II* are summarized in Table III.

TABLE III. SUMMARY OF FACTS WHICH PERTAIN TO IMPORTANT CHARACTERISTICS OF TEACHER-MADE OBJECTIVE TESTS

Type of Item	Number of Items	Per Cent of All Items
Completion	698	28.6
Matching	363	14.9
Multiple-Choice	371	15.2
True-False	973	39.9
Location	18	.7
Unclassified	16	.6
Total	2439	99.9

Number of tests covering designated text material	42
Number of tests covering semester of work	21
Total number of tests	63

The classification of items used in Table III follows in general the classification presented by Ruch in his *Objective or New-Type Examination.*[3] In the case of the four types most frequently used, a number of variations appeared in the tests; that is, there were a number of sub-types of each of these main types. The data presented in Table III show that approximately 40 per cent of all the items

[1] The relation between the length of a test and its reliability is not so simple and direct as has often been implied in the measurement literature. Recently the fact that the reliability of a test may be increased by *decreasing* its length has been demonstrated. This means that the types of item which are added to or eliminated from a test will determine the resulting effect upon reliability and not the increasing or decreasing of the number of items per se. For an illuminating discussion of this problem see R. R. Willoughby, "The Concept of Reliability," *Psychological Review,* XLII (March, 1935), 153-165. Also, W. A. Brownell, "On the Accuracy with Which Reliability May Be Measured by Calculating Test Halves," *Journal of Experimental Education,* I (March, 1933), 204-215.

[2] An asterisk following the number of a table indicates that the table appears in the Appendix of this report.

[3] Ruch, *op. cit.,* pp. 189-190.

were some form of the true-false type. The second most frequently used type of item was the completion;[4] about one-fourth of all the items are classified under this heading. Finally, approximately 15 per cent of the items were of the matching type and about the same number of the multiple-choice type. Table II* reveals, also, that forty-four, or 69.8 per cent, of all the tests were composed of two or more types of item.

Two attacks made. Data which bear upon the teacher-made test part of the investigation were gathered by two methods which differed in minor respects. In the case of both methods the basic technique already described was used. The chief difference consisted in the fact that in one attack the tests were constructed to cover an exactly designated unit of text material, while in the other the tests were made to cover the work of a semester. In the latter case the same texts were used by all teachers whose tests were paired, but no attempt was made to guarantee that the tests covered identical text material. Since the teachers whose tests were paired taught in the same or closely affiliated school systems and followed the same course of study, the differences, in reality, in text matter were relatively small.

DETAILED DESCRIPTION AND ANALYSIS OF FIRST ATTACK

Subject matter covered by tests. The teachers who taught a given subject at a given grade level[5] were asked to agree upon a body of text material on which the tests were to be based. The text matter varied from grade to grade and from subject to subject, but was identical for any given subject and any grade level.

Instructions for making tests. As stated under the topic "Example of technique," the teachers were instructed to make tests that they considered adequate measuring instruments for determining the achievement of their pupils with reference to the designated text material. The tests were to be designed to furnish an adequate basis for giving the pupils marks which accurately represented their achievement. All of the teachers habitually made use of the new-type test in their regular work; but in order to insure that the tests be of the type desired, each teacher was given a mimeographed list of examples of the more frequently used types of objective test.

[4] The term "completion" as used here is synonymous with Ruch's "Recall types."

[5] Fifth and sixth grade geography, fifth and sixth grade and high-school history, and sixth grade health were the subjects and grade levels used in the investigation.

However, the teachers understood that this list was suggestive and not limiting.

In addition to the two restricting suggestions mentioned on page 28, the following points were emphasized:

1. Make such directions for the test as you would consider adequate for an effective administration of your test.
2. Use any type of item or as many types of item as you wish to use in making your test.
3. Let the length, as well as all other characteristics, of the test be determined by your judgment as to the number of items necessary in order that the test be an adequate measuring instrument.
4. Do not confer regarding the nature and content of your test with other teachers who are taking part in the investigation.

Pairing teacher-made tests. Each test was mimeographed in exactly the form suggested by the teacher who constructed the test. In order that the tests might be easily identified each teacher was assigned a code number which was placed on the first page of the test. Thus a test in history might have the code number IVH5 in the upper left-hand corner of the first page of the test, which identified the test as having been made by teacher number IV in fifth grade history.

Teachers whose tests were made to cover the same text matter were paired for administration of the tests. Since each group of pupils was to take two tests (the test made by the teacher of the group and a second test constructed by another teacher), a second test had to be chosen for each group of pupils. In order to avoid any tendency to pair tests which were especially similar or especially dissimilar in content or form, teachers were paired by the chance method, without regard to the nature of the tests which they constructed. After the second test for a group had been thus selected, sufficient mimeographed copies of the two tests were placed in the hands of the teacher of the group for administration.

Directions for administering the tests. After the plan for administering the tests had been carefully explained during an interview, each teacher received a mimeographed set of directions designed to neutralize the effect of practice and fatigue in the pupils who took the tests. A copy of the directions follows:

Directions for Giving History and Geography Tests

NOTE: It is extremely important that every teacher follow the directions exactly. It is only by following the directions carefully that the results will be of value.

1. In order that the results on the two tests you are giving may be comparable it is necessary that the tests be given at the same time. If one test is given before the other there will be some practice effect or perhaps some confusion. You are requested, therefore, to give the tests in the following manner. Let us call your own test Test A, and the other test which you are giving Test B. Give every other row in your classroom Test A, and the remaining children Test B. Then during the second half of the period give Test A to the children who previously had taken Test B, and Test B to the children who had taken Test A. For example, children occupying rows 1, 3, 5, and 7 would take Test A while the children in rows 2, 4 and 6 are taking Test B. Then during the second half-period, children occupying rows 1, 3, 5 and 7 would take Test B while those occupying rows 2, 4, and 6 take Test A. This procedure, as you see, will tend to equalize practice effect or any confusion that a test might cause in a child's mind.

2. Write down any directions that you give to the pupils other than those appearing on the test, both for your own test and for the other test you give. Give such directions as you consider necessary but remember to keep an accurate account of the directions that you give. (Use the attached sheet for reporting this information.)[6]

The pupils were allowed sufficient time to complete the tests. Since the pupils looked upon these tests as their regular periodic examination, the motivation was relatively strong. The pupils were not acquainted with the fact that they were taking a test made by another teacher. In order that the situation might be as normal as possible, in each case the two tests were given by the teacher of the group that took the tests.

Scoring of the tests. At the same time that he constructed his test, each teacher made out a key and a set of directions for scoring his test. The teachers scored their own tests; the second tests were scored by the investigator with the use of the proper key. Precautions were taken in order to avoid errors in scoring.

Teachers, tests, pupils, and test papers involved. The scope of this part of the study can be best understood when one is acquainted with facts relative to the number of teachers, of tests, of pupils, and of test papers which were involved in gathering the data. Table IV is presented in order to supply these facts.

Table IV shows that twenty-four teachers[7] assisted in this study. These teachers constructed a total of forty-two tests which were ad-

[6] Other directions of a routine nature which pertained to such matters as date on which tests should be administered were included on the mimeographed sheet of directions, but these points, not bearing upon the issue under discussion, are not given here.

[7] In a number of cases the same teacher made two tests; for example, a fifth grade teacher would make a test in fifth grade history and in fifth grade geog-

TABLE IV. TEACHERS, TESTS, PUPILS, AND TEST PAPERS INVOLVED IN THE STUDY OF TEACHER-MADE OBJECTIVE TESTS WHICH COVER SHORTER UNITS OF MATERIAL

Classification	Frequency
Teachers	24
Tests	42
Pupils*	2,255*
Test papers	4,510

*When duplications are subtracted the number of pupils is 1,575.

ministered to 2,255 pupils.[8] Since each child took two tests, the data of this phase of the investigation are based upon 4,510 individual test papers.

Variation of the first attack. The general techniques of the first attack were altered somewhat in the case of two groups in high-school history and in the case of the health group. The health tests were given to a group of one hundred pupils. These tests were constructed and administered in accordance with the directions already described except that the paired tests were not given to the pupils of the teachers who made the tests. This permitted significant comparisons of the two tests as measuring instruments.

The same procedure was followed in the case of two pairings of high school history tests. These four tests or two pairs of tests may be identified in the subsequent chapters by the code numbers ISHSV-IISHSV and IIISHSV-IVSHSV.

DETAILED DESCRIPTION AND ANALYSIS OF SECOND ATTACK

Unit for testing. As has been stated already, the unit over which the tests were made in the second attack was a semester of work in a given subject. Teachers were paired only when they used the same text. The point has been made that since the texts were the same, the school systems the same or closely related, and the general courses of study were identical, the tests in reality covered very similar subject matter. The tests were the regular semester final examinations except that the two tests were given instead of the

raphy. Or in the case of those schools which use the platoon system, the geography teacher would construct tests for fifth and sixth grade geography. The number of teachers given above represents the number of individual teachers who assisted in the study.

[8] In many instances the same pupils took a geography test and a history test. For example, a fifth grade pupil would take a fifth grade history test and a fifth grade geography test. In calculating the number of pupils these pupils were counted twice. The total number of pupils when such duplication is eliminated is 1,575.

one which would have been given normally. With the exception mentioned in the preceding paragraph the procedures used in the second attack were identical with those used in the first. The data based upon semester tests were gathered in seventh grade geography, seventh grade history, and high-school history.

Teachers, tests, pupils, and test papers involved. Table V shows the number of teachers, tests, pupils, and test papers which were involved in the second attack.

TABLE V. TEACHERS, TESTS, PUPILS, AND TEST PAPERS INVOLVED IN STUDY OF TEACHER-MADE OBJECTIVE TESTS WHICH COVER A SEMESTER OF WORK

Classification	Frequency
Teachers	11
Tests	21
Pupils*	878*
Test papers	1,756

*When duplications are subtracted the number of pupils is 516.

The data in Table V show the scope of the second attack upon the teacher-made test phase of the research. The table shows that eleven teachers constructed a total of twenty-one semester tests; that these tests were given to 878 pupils; and that since each pupil took two of the tests there was an aggregate of 1,756 test papers.

TYPES OF RESULTS SECURED

The results of the part of this research which is based upon teacher-made tests are presented and interpreted in Chapters IV, V, and VI. Table I (see p. 29) was presented as an illustration of the basic technique used in this part of the investigation. The two sets of percentage scores for each child or each group may be considered as the basic data. How should these data be analyzed in order to show most clearly the extent of disparity in the results from teacher-made objective tests? Three methods of analysis are employed: correlation, percentile point disparity, and teacher-mark disparity.

As an illustration of the types of analysis which are made, the facts presented in Table I are given again at this point and to these facts are added data basic to the analyses which are presented in Chapters IV, V, and VI. A sample set of these data appears as Table VI. The paired tests in this instance were Test IH5 and Test VH5. The tests were given to the pupils listed in column 1 of Table VI. Column 2 shows the score that each child made on

Test VH5, and column 3 shows the corresponding score for Test IH5.

a. The first analysis consists of correlating the two sets of scores in columns 2 and 3. The coefficient in this case is .239.

b. The second type of analysis is made in terms of percentile rank disparity. The figures in column 4 show the percentile rank of each pupil on Test VH5, and those in column 5 reveal his corresponding percentile rank for Test IH5. These facts permit the second type of analysis. For example, Pupil 1's percentile point position on Test VH5 was 70 (i. e., 70 per cent of the pupils made scores lower than Pupil 1), whereas his percentile position on Test IH5 was 83. Thus there was a difference of 13 in Pupil 1's percentile rank on the two tests. The second analysis is made in terms of these differences.

c. Finally, the letters in columns 6 and 7 represent the marks which the child's performance on the two tests would warrant according to the system of marking used in the school system in which

TABLE VI. SAMPLE SET OF DATA ILLUSTRATING TYPES OF ANALYSES USED IN THE STUDY OF TEACHER-MADE OBJECTIVE TESTS

Pupil (1)	Score on Test VH5 (2)	Score on Test IH5 (3)	P. R. on Test VH5 (4)	P. R. on Test IH5 (5)	Mark on Test VH5 (6)	Mark on Test IH5 (7)
1	90	79	70	83	A	C
2	85	71	58	45	B	C
3	70	87	25	96	C	B
4	92	75	81	64	A	C
5	95	79	93	83	A	C
6	67	71	20	45	F	C
7	92	75	81	64	A	C
8	60	67	13	32	F	F
9	37	54	1	4	F	F
10	92	71	81	45	A	C
11	80	71	49	45	B	C
12	92	79	81	83	A	C
13	72	75	35	64	C	C
14	97	67	99	32	A	F
15	72	66	35	27	C	F
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
40	85	62	58	18	B	F
41	75	71	42	45	C	C
42	60	75	13	64	F	C

N.....42
Mean P. R. Difference.....26.6
Mean Mark Interval Difference..... 1.29
r..... .239

the study was made. As an illustration, Pupil 1's mark on Test VH5 was *A*, and his mark on Test IH5 was *C*. This means that the two marks differed to the extent of two mark intervals. These facts permit an analysis in terms of teacher-mark disparity or variation, the third type of analysis used. These analyses of results together with some interpretation of their meaning appear in the next three chapters.

CHAPTER IV

Correlation Analysis of Results from Paired Informal or Teacher-Made Objective Tests

Purpose of chapter. In this chapter the data secured as described in Chapter III are analyzed by the use of the correlation technique. Each of sixty-eight groups of pupils was given tests which were constructed by two different teachers to measure the same abilities, and the two sets of scores which resulted were correlated. The purpose of this chapter is to present and interpret the coefficients of correlation thus secured.

The disparity between paired teacher-made tests is expressed in this study in three ways or in terms of three types of indexes. One index of the extent of disparity is the degree to which scores on paired tests are related as represented by correlation.[1] The correlation coefficients presented in this chapter are offered, therefore, as one index to the extent of variability in teacher-made objective test results.

Correlation between tests which cover identical units of text material

Fifth grade geography. In fifth grade geography paired tests were administered to ten groups of pupils. As a rule, for purposes of analysis, one teacher's pupils were considered as a test group.[2] The scores made by the pupils in each test group on one test were correlated with the scores made by the same pupils on the second test. The ten coefficients secured are presented in Table VII.

Table VII shows that the mean correlation for the ten groups is .523. The highest correlation revealed is .71 and the lowest is .247.

The fact should be kept in mind that the test groups used in fifth and sixth grade geography and history included all the pupils in these grades in a particular school. In cases where N is more than forty-five the test group was composed of high and low ability sections of pupils. The correlation coefficients were calculated from the test scores of both sections taken together. This means that the

[1] When applied to tests, coefficients of correlation secured in this manner are frequently designated as coefficients of reliability. See pp. 69-70 for a brief discussion of the meaning of correlation coefficients as applied to tests.

[2] In all cases, the results are presented in terms of test groups.

TABLE VII. FIFTH GRADE GEOGRAPHY—CORRELATION BETWEEN INFORMAL* OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
XIIG5 –VG5†	.710	33
IIG5 –XIIIG5	.662	55
XIIIG5–IIG5	.659	86
VG5 –XIIG5	.643	42
VIIIG5 –VIIG5	.568	99
IG5 –IXG5	.523	72
IXG5 –IG5	.452	65
VIIG5 –VIIIG5	.438	44
IIIG5 –IVG5	.344	28
IVG5 –IIIG5	.247	88
Mean	.523	61.2

*The terms "informal" and "teacher-made" are used as synonyms.
†Teacher XII, Geography, Grade 5 — Teacher V, Geography, Grade 5.

relationship would probably have been somewhat lower if the r's had been based upon the separate sections, for in such case the spread of the scores would have been less.

Sixth grade geography. Table VIII presents the same facts for sixth grade geography as were given for fifth grade geography in Table VII.

TABLE VIII. SIXTH GRADE GEOGRAPHY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
XIG6 –VIG6	.770	37
VIG6 –XIG6	.718	45
IIG6 –IVG6	.621	78
IXG6 –XIIIG6	.539	66
XIIIG6–IXG6	.395	67
VIIIG6 –XG6	.320	84
IVG6 –IIG6	.197	71
Mean	.508	64

The range of coefficients for sixth grade geography is from .197 to .770 and the average correlation for all test groups is .508.

Summary for identical unit tests in geography. In fifth grade geography and sixth grade geography a total of seventeen paired tests were given and the scores for each test group correlated. The mean correlation for these groups is approximately .50.

Fifth grade history. Eight groups of children were given paired tests in fifth grade history. When the test scores were correlated, the coefficients which appear in Table IX were found.

TABLE IX. FIFTH GRADE HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
IH5 –VH5	.684	54
XIIH5 –IVH5	.569	32
VIIH5 –IIH5	.550	44
IXH5 –VIIIH5	.532	37
IIIH5 –VIIH5	.438	28
IVH5 –XIIH5	.382	75
VIIIH5 –IXH5	.352	64
VH5 –IH5	.239	42
Mean	.468	47

The range of the coefficients is approximately the same as in the case of fifth and sixth grade geography. As is shown in Table IX the mean correlation for fifth grade history is .468.

Sixth grade history. The relationship between paired tests in sixth grade history is shown in Table X.

Table X reveals a mean correlation of .535 for the nine test groups. The highest coefficient is .661 and the lowest is .377.

TABLE X. SIXTH GRADE HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
IIH6 –VIH6	.661	31
XIH6 –IVH6	.633	37
VIIIH6 –IH6	.606	65
VIH6 –IIH6	.591	45
XH6 –IIH6	.550	41
IH6 –VH6	.537	74
IVH6 –XIIIH6	.457	44
IVH6 –XIH6	.407	41
XIIIH6–IVH6	.377	64
Mean	.535	49.1

High-school history. When the scores from paired tests in high-school history were correlated, the coefficients presented in Table XI were found.

The relationship between the paired tests which covered a short unit in high-school history was the highest found in the teacher-made test phase of the study. The mean r in this case is .711. The relatively high degree of relationship shown in Table XI may, in part at least, be accounted for by two facts: first, of the four teachers who constructed the tests used, three taught in the same high school and had frequent conferences concerning testing procedures.

TABLE XI. HIGH-SCHOOL HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Tests	r	N
IIIHSH-IHSH	.837	32
IIIHSH-IVHSH	.820	32
IVHSH-IHSH	.729	33
IIHSH -IHSH	.725	34
IIHSH -IVHSH	.669	32
IIHSH -IVHSH	.486	34
Mean	.711	32.8

Although the teachers did not confer when they were constructing the tests used in this study, the unusually close contact between the teachers very likely decreased the variability which results from the selection of items for a test. In the second place, the testing unit was relatively short, whereas the tests were long and specifically factual in nature.

Variation in technique for high-school history. As a further check upon the amount of disparity between objective tests in high-school history the four high-school history tests were paired and given to groups of pupils who were not the pupils of either of the teachers who constructed the tests used.[3] Tests IHSHVI and IIHSHVI were considered as one pair and were given to a group of thirty-six pupils. Tests IIIHSHV2 and IVHSHV2 constituted a second pairing and were administered to forty-four pupils. When the scores were correlated the following coefficients were obtained:

Tests	*r*
IHSHVI-IIHSHV1	.654
IIIHSHV2-IVHSHV2	.341

Reference to Table XI shows that the first of the two pairings just listed correlated .725 when the two tests were administered according to the regular procedures as compared with .654 when the variation method was used. Even more striking is the fact that the second variation pairing showed a correlation of .820 (Table XI) when administered according to the regular procedure, as compared with a correlation of .341 when the method was varied. These facts seem to constitute some evidence that the relatively high relationship revealed in Table XI was due in some degree to factors other than those in the tests themselves.

Sixth grade health tests. A pair of health tests was administered

[3] See p. 34 for a detailed description of the procedure.

to a group of one hundred pupils in the sixth grade. The correlation between the two sets of scores was .402.

Summary for all identical text material tests. Table XII presents a summary of the forty-three coefficients of correlation which resulted from the correlating of paired tests constructed to cover identical units of text matter.

Table XII reveals that paired tests were administered in history in the fifth and sixth grades and in high school; in geography in the fifth and sixth grades; and in health at the sixth grade level. In column 3 the mean r for each subject-grade group is given. Of the eight mean coefficients presented, six are below .55. The mean of the mean coefficients as shown in Table XII is .51. This summary mean is based upon the correlation between the scores from forty-three test groups which involve a total of 2,255 pairs of individual test papers.

TABLE XII. SUMMARY OF CORRELATION BETWEEN TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Subject	Grade	Mean r	Mean N	Number of Test Groups
Geography	5	.523	61	10
Geography	6	.508	64	7
History	5	.468	47	8
History	6	.535	49	9
History	S.H.S.	.711	32	6
History V1	S.H.S.	.654	44	1
History V2	S.H.S.	.341	36	1
Health	6	.402	100	1
Summary mean		.517	54	

CORRELATION BETWEEN TESTS CONSTRUCTED TO MEASURE A SEMESTER OF WORK

The results already presented show that the relationship between paired objective tests which were constructed to measure identical bodies of text matter expressed in terms of a coefficient of correlation is approximately .50. The question arose as to what the relation would be if semester examinations were used instead of tests constructed to cover shorter units of work. In order to illuminate this problem, paired semester tests were administered to twenty-five test groups, and the two sets of scores secured for each group were correlated.

Seventh grade geography. Ten test groups were given paired

semester tests in seventh grade geography. The correlations found are shown in Table XIII.

The range of the coefficients is relatively large in this case. The highest correlation is approximately .85, whereas the lowest shows a slight negative relationship or a coefficient of –.21. The mean coefficient is .413.

TABLE XIII. SEVENTH GRADE GEOGRAPHY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER A SEMESTER OF WORK

Tests	r	N
IIG7 –IG7	.845	39
IG7 –IIG7	.657	43
IIG7 –IG7	.632	40
IIG7 –VG7	.594	40
IG7 –VG7	.581	40
IG7 –IVG7	.393	25
IG7 –IIIG7	.284	37
IIG7 –IIIG7	.246	34
IVG7–VG7	.117	28
VG7 –IIG7	—.212	36
Mean	.413	36.2

Seventh grade history. Table XIV presents the corresponding facts for seventh grade history.

TABLE XIV. SEVENTH GRADE HISTORY—CORRELATION BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER A SEMESTER OF WORK

Tests	r	N
IIH7e – IIIH7	.590	34
IIIH7 – IIH7e	.643	27
IIH7e – IH7c	.642	32
IIH7c – IH7c	.616	40
IVH7 – VH7	.600	28
IH7a – IVH7	.589	39
VH7 – IIIH7	.465	36
VIH7 – VH7	.398	31
IIH7a – IVH7	.327	44
IVH7 – IIH7a	.309	40
IIH7b – IIIH7	.078	36
IH7e – IIH7c	.053	35
IIH7d – VH7	—.092	36
Mean	.401	35.2

Tests were administered to thirteen test groups in seventh grade history. The mean of the thirteen coefficients is .401 and the median is .465.

High-school history. Only two pairs of semester tests were given in high-school history, but the two coefficients secured are of inter-

est, because the fact that the relationship is quite similar in magnitude to that found in other phases of the study supports the general results. The four semester tests were paired as follows:

1. IHSB-IIHSB
2. IHSF-IIHSF

Correlation of the scores from the first pair of tests (No. 1) revealed an r of .276, and of the second pair of tests (No. 2) an r of .420. The mean coefficient for the high-school semester tests is approximately .35. This relatively low relationship becomes more interesting when one recalls that the four tests used were perhaps the most detailed and carefully constructed of the semester tests used in this study. Reference to Table II* reveals the fact that the four tests were composed of the following numbers of items:

Test	*No. of items*
IHSB	105
IIHSB	65
IHSF	100
IIHSF	80

Thus the lack of relationship revealed can hardly be attributed to the length of the tests.

Summary for all semester tests. Paired semester tests were given to twenty-five test groups, ten in seventh grade geography, thirteen in seventh grade history and two in high-school history. The following mean r's were obtained:

Subject and grade	*Mean r*
Seventh grade geography	.413
Seventh grade history	.401
High-school history	.348

The mean of these three means is .387. That is, the average relationship between twenty-five paired semester tests is approximately represented by a coefficient of .40. It is of interest to recall at this point that the comparable mean for the shorter unit tests was approximately .50.

CORRELATION BETWEEN PAIRED TESTS IN GEOGRAPHY

Table XV shows the correlation between all (semester and shorter unit) tests in geography.

Paired tests were given to groups of pupils in fifth, sixth, and seventh grade geography. Table XV shows the mean correlation for each of these grades and the mean of these means. The summary mean which is based upon twenty-seven coefficients is .481.

TABLE XV. MEAN CORRELATION FOR ALL PAIRED TESTS IN GEOGRAPHY

Subject	Grade	Mean r	Mean N
Geography	5	.523	61.2
Geography	6	.508	64.0
Geography	7	.413	36.2
Mean of means		.481	53.8

CORRELATION BETWEEN PAIRED TESTS IN HISTORY

How does the relationship between the geography tests compare with that between history tests? The facts presented in Table XVI answer this question.

TABLE XVI. MEAN CORRELATION FOR ALL PAIRED TESTS IN HISTORY

Subject	Grade	Mean r	Mean N
History	5	.468	47.0
History	6	.535	49.1
History	7	.401	36.2
History	S.H.S.	.515	33.9
Mean of means		.479	41.3

The mean of the means for all the paired tests in history (forty test groups) is .479. The fact that this figure is strikingly similar in magnitude to that obtained for all geography tests should be noted, for this similarity of the two summary means is evidence that the relationship found approximates the relationship which actually exists between informal objective tests.

RESULTS FROM ALL TESTS

Sixty-eight pairs of tests were given in connection with the study of teacher-made objective tests, twenty-seven pairs of tests in geography, forty pairs in history, and one pair in health. Table XVII shows the mean correlation for each subject-grade group and the average of these means.

Table XVII indicates that the mean of the eight means which are based upon sixty-eight coefficients is .470.[4] The most convincing feature of Table XVII is the similarity in magnitude of the eight mean coefficients. No mean coefficient is less than .40 and none is above .54. This consistency of the mean r's which are shown in

[4] The writer is aware that there has been criticism of averaging coefficients of correlation. However, one should keep in mind that the purpose of computing this mean is to indicate a tendency in the extent of relationship which is found when a number of correlations are considered.

TABLE XVII. SUMMARY OF CORRELATIONS BETWEEN PAIRED TESTS FOR ALL SUBJECTS AND ALL GRADE LEVELS

Subject	Grade	Mean r	Mean N	Number Test Groups
Geography	5	.523	61.2	10
Geography	6	.508	64.0	7
Geography	7	.413	36.2	10
History	5	.468	47.0	8
History	6	.535	49.1	9
History	7	.401	35.2	13
History	S.H.S.	.515	33.9	10
Health	6	.402	100.0	1
Mean of means		.470	53.3	

Table XVII seems to warrant the following tentative generalization: The correlation between teacher-made objective tests constructed by different teachers to measure pupil acquaintance with the same subject matter and scored objectively is approximately .50.

TRADITIONAL AND OBJECTIVE TYPE TESTS COMPARED

A study of traditional (essay) type tests, very similar to the study of objective type tests reported in this chapter, was made by Monroe and Souders.[5] The availability of these data makes possible an interesting comparison. Table XVIII presents this comparison. The first half of the table is a summary distribution of the results from Monroe's and Souders' study of paired traditional type tests. Only those data directly comparable with the findings of the present study (of new-type tests) are presented. The second half of Table XVIII is a summary distribution of the coefficients found when scores from paired objective tests were correlated in the present study.

The median correlation for the traditional type tests (thirty-four coefficients) is .65, whereas the corresponding median for the objective type tests (sixty-eight coefficients) is .54.

The difference between the two medians is not great, and therefore it is not wise perhaps to conclude from these facts that objective tests show more variation than traditional type tests; however, the data presented in Table XVIII warrant the conclusion that objective tests as used by classroom teachers are probably no more free from factors which cause variation as measured by the methods used in this study than are traditional type tests. If one

[5] Monroe and Souders, *op. cit.*, p. 77. Monroe's and Souders' data are quoted by Ruch, *op. cit.*, chap. iii, in connection with his discussion of "objections to traditional type tests."

TABLE XVIII. COMPARISON OF DISTRIBUTIONS OF CORRELATIONS FOR TRADITIONAL TESTS AND FOR OBJECTIVE OR NEW-TYPE TESTS

Size of r	Traditional Type Test* – Frequency	Objective Type Test – Frequency
.95	0	0
.90	0	0
.85	2	0
.80	2	3
.75	4	1
.70	4	4
.65	6	7
.60	4	9
.55	2	9
.50	2	4
.45	3	4
.40	1	5
.35	0	6
.30	1	5
.25	1	2
.20	0	3
.15	0	1
.10	0	1
.05	1	2
.00	0	0
—.05	0	1
—.10	0	0
—.15	1	0
—.20	1	0
—.25	0	1
Total	34	68
Median	.65	.54

*Data adapted from Monroe and Souders, *op. cit.*, pp. 32, 33.

may judge by the literature, the general opinion of writers on educational measurements has been contrary to this conclusion.[6] Although no facts have been available on this point, the assumption that the new-type or objective test as a measuring instrument is less subjective or variable in general than the essay type test seems to have been widely accepted. The findings of this investigation are at least a beginning toward the substitution of relevant facts for unsubstantiated opinion.

Further, the facts given in Table XVIII seem to indicate that objectivity in scoring (the distinguishing mark of the objective test) as a feature of testing does not decrease significantly the disparity in the results from such tests. That is to say, objectivity in scoring seems to be a relatively unimportant factor in determining the extent of variation which grows out of the measurement situation as a whole.

[6] See E. V. Pullias, "A Study of Current Opinion Concerning Objective Tests," *Educational Method,* XVI (April, 1937), 348-356.

CHAPTER V

Percentile Point Analysis of Results from Paired Informal Objective Tests

Purposes of chapter. In Chapter IV the data based on paired teacher-made objective tests were analyzed by the correlation technique. In this chapter the same basic data are analyzed by what may be called the *percentile rank disparity* technique. This method has been described and illustrated on p. 36. Briefly stated, the procedure consists of comparing the percentile rank position of a pupil on one test with his corresponding position on a second test when both tests were constructed to measure knowledge of identical material. The purpose of the percentile rank analysis may be clarified by illustration. If a group of children were weighed on two different scales and percentile rankings determined for each series of weighings, one would expect relatively little disparity between the two rankings. The child who ranked heaviest as weighed by one set of scales would, within limits, rank heaviest as weighed by the second set of scales; therefore, the differences in percentile ranks would approach zero. If objective tests are used as measuring instruments what will be the extent of the percentile rank differences? To answer the question just stated is the central purpose of this chapter.

Tests which cover identical units of text matter

Fifth grade geography. Paired tests were given to ten groups of pupils in fifth grade geography. In order to illustrate the technique used the percentile rank disparity found between these tests is shown in Table XIX.

One should remember when reading Table XIX that the step intervals are percentile rank differences. For example, the *1* in the frequency column opposite 80 – indicates that the difference between this pupil's percentile ranks on the two tests which he took was approximately 82.5 (midway in the interval, 80-84). He had a percentile rank of ten on one test and ninety-one on another. The cumulative frequency column shows the cumulated frequencies reading downward on the step interval scale. For example, sixteen of the pupils were 65 or more percentile points apart on the paired tests. The cumulative percentage distribution is calculated in the same gen-

TABLE XIX. FIFTH GRADE GEOGRAPHY—DISPARITY IN TERMS OF PERCENTILE POINTS BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—	0	0	0	0
90—	0	0	0	0
85—	0	0	0	0
80—	1	1	0.2	0.2
75—	1	2	0.2	0.4
70—	6	8	1.0	1.4
65—	8	16	1.3	2.7
60—	11	27	1.8	4.5
55—	9	36	1.5	6.0
50—	17	53	2.8	8.8
45—	19	72	3.1	11.9
40—	20	92	3.3	15.2
35—	33	125	5.4	20.6
30—	36	161	5.9	26.5
25—	54	215	8.8	35.3
20—	62	277	10.1	45.4
15—	73	350	11.9	57.3
10—	71	421	11.6	68.9
5—	99	520	16.1	85.0
0—	92	612	15.0	100.0
N	612		100.0	
M	21.5			

eral manner as the cumulative frequency column. As an illustration, in the case of 8.8 per cent of the pupils their percentile ranks showed a disparity of 50 or more percentile points.

Table XIX shows that the mean percentile rank disparity for the ten test groups is 21.5. One hundred and twenty-five pupils, or approximately 20 per cent of the group, showed a percentile rank difference of 35 or greater. More than one-third (35 per cent) of the group revealed a minimum rank difference of 25 percentile points.

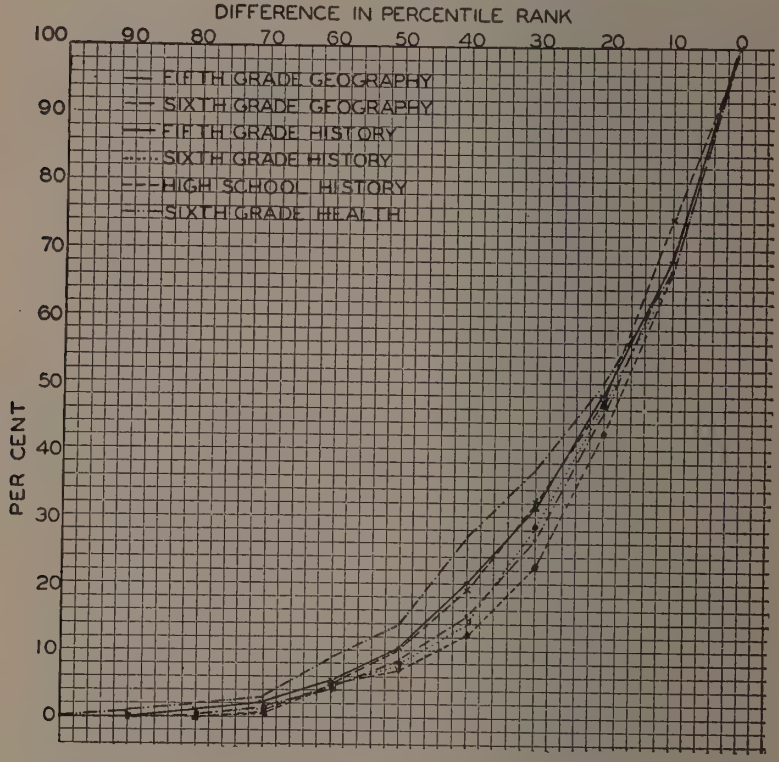
Other identical unit tests. Data[1] similar in nature to those presented in the foregoing section were obtained for sixth grade geography, fifth and sixth grade history, high-school history, and sixth grade health. The consistency of the extent of percentile rank disparity is of interest. One will recall that approximately one-third (35 per cent) of the pupils in fifth grade geography were 25 or more percentile points apart on the two tests. The corresponding figures for the remaining subject-grade groups are as follows: sixth grade geography, 37 per cent; fifth grade history, 39 per cent; sixth grade

[1] Complete data for each subject is not presented because of space limitations. Anyone desiring such facts should correspond with the author.

history, 36 per cent; high-school history, 29 per cent; sixth grade health, 42 per cent.

The consistency of the distributions is shown even more clearly in FIG. I. Note how nearly the curve for each subject-grade level represented in FIG. I approximates at all points the curves for the other subject-grades. As an example, in the case of all subject-grade levels approximately 50 per cent of the pupils showed a percentile point disparity of 20 or more.

FIGURE I. PERCENTILE CURVE FOR IDENTICAL UNIT TESTS



Summary for all identical text material tests. The percentile rank disparity revealed when forty-three groups of pupils (2,255 individuals) took paired tests constructed to cover identical text material is shown in Tables XX and XXI.

Table XX presents a composite distribution based upon the forty-three group distributions. The range of percentile rank disparity as shown in Table XX is from 0 to 99. There are cases in every in-

TABLE XX. SUMMARY DISTRIBUTION—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—	1	1	0.04	0.04
90—	1	2	0.04	0.08
85—	3	5	0.13	0.21
80—	3	8	0.13	0.34
75—	8	16	0.35	0.69
70—	16	32	0.71	1.40
65—	28	60	1.24	2.64
60—	45	105	2.00	4.64
55—	36	141	1.60	6.24
50—	64	205	2.84	9.08
45—	70	275	3.10	12.18
40—	99	374	4.39	16.57
35—	126	500	5.59	22.16
30—	144	644	6.39	28.55
25—	170	814	7.54	36.09
20—	220	1034	9.76	45.85
15—	266	1300	11.80	57.65
10—	249	1549	11.04	68.69
5—	369	1918	16.36	85.05
0—	337	2255	14.95	100.00
N	2255		100.00	
M	22.2			

terval from 0-4 to 95-99. The mean disparity is 22.2. Table XX shows that 205 pupils, or approximately 10 per cent of the group, had percentile rankings on one of the tests taken which varied 50 or more percentile points from their rankings on the second test.

Five hundred pupils, or somewhat more than one-fifth of the group, showed a disparity of 35 or more percentile points; and distinctly more than one-third of the total group revealed a disparity ranging from 25 percentile points upward.

TABLE XXI. SUMMARY OF MEANS—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

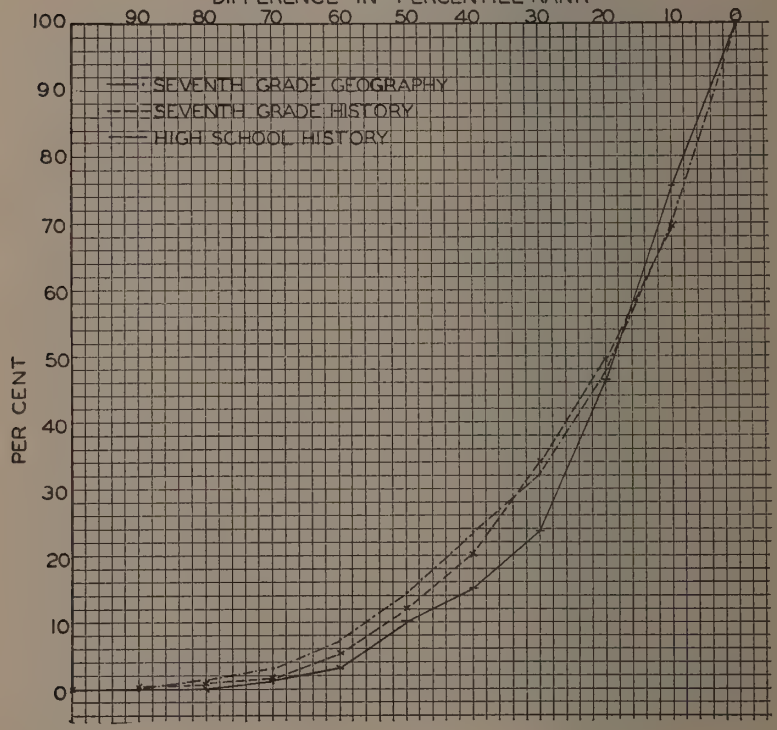
Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography	5	21.5	61.2	10
Geography	6	22.9	64.0	7
History	5	23.0	47.0	8
History	6	21.6	49.1	9
History	SHS	17.7	32.8	6
History V	SHS	23.3	40.0	2
Health	6	25.8	100.0	1
Mean of means		22.2	56.3	

Table XXI is a more general summary of the facts presented in Table XX. The mean percentile rank disparity for each subject and grade level is shown. The greatest mean disparity found was 25.8 (health), and the least was 17.7 (high-school history). The remaining five means are very similar in size, clustering very closely around the mean of the means. The consistency in magnitude of these means seems to indicate that they are relatively reliable, that is, that considerable confidence may be put in a generalization based upon these averages. The summary (based upon forty-three test groups) is 22.2 percentile points.

TESTS CONSTRUCTED TO MEASURE A SEMESTER OF WORK

The findings reported in the foregoing part of this chapter show that the mean amount of disparity in terms of percentile rank differences between shorter unit tests is approximately 22. Since the purpose of the investigation was to determine the amount of dis-

FIGURE II. PERCENTILE CURVE FOR SEMESTER TESTS



parity between objective tests as they are generally used in school situations, the next step consisted of extending the study to semester examinations.

How does the disparity between paired semester tests compare with that shown to exist between tests which cover shorter units of work? In order to answer this question, paired semester tests were administered to twenty-five test groups (878 individuals) and the amount of percentile rank disparity was calculated.

Seventh grade geography, seventh grade history, and high-school history. The percentile rank disparity for groups of pupils in seventh grade geography, in seventh grade history, and in high-school history was determined. These distributions are very similar to those for shorter unit tests. On the whole the disparity was slightly greater in the case of semester tests. The consistency of the semester test distributions is shown in Fig. II. The three curves tend to be very similar. As an illustration, the point 20 on the percentile rank difference scale in each case corresponds approximately to the point 50 on the percentage scale; that is, in each of the three groups represented by the curves in Fig. II, about 50 per cent of the frequencies appear at or above the point 20 on the percentile rank difference scale. (See Fig. I for consistency between the two sets of curves.)

Summary for all semester tests. Summaries of the disparity between all paired semester tests are presented in Tables XXII and XXIII. Table XXII is a summary distribution based upon twenty-five distributions for smaller groups.

Taken as a whole, the semester examinations revealed very nearly the same amount of disparity as the tests which covered shorter units of work. The mean for the semester tests is 23.9, as compared with 22.2 for the shorter unit tests. The distribution of frequencies is very similar in both cases. (Compare cumulative percentage distributions in Table XX and Table XXII.)

Table XXIII gives the means for the three groups of semester tests.

The chief purpose of bringing these means together is to emphasize their consistency. Reference to Table XXIII indicates that the range is slightly less than 3 percentile points. The data gathered from the study of semester tests seem to warrant the conclusion that these tests show practically the same amount of disparity as tests constructed to cover shorter units of work.

TABLE XXII. SUMMARY DISTRIBUTION—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER A SEMESTER OF WORK

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—	0	0	0	0
90—	2	2	0.23	0.23
85—	2	4	0.23	0.46
80—	5	9	0.57	1.03
75—	5	14	0.57	1.60
70—	7	21	0.80	2.40
65—	15	36	1.70	4.10
60—	20	56	2.28	6.38
55—	19	75	2.16	8.54
50—	39	114	4.44	12.98
45—	31	145	3.53	16.51
40—	43	188	4.90	21.41
35—	39	227	4.44	25.85
30—	60	287	6.83	31.68
25—	60	347	6.83	38.51
20—	81	428	9.23	47.74
15—	85	513	9.68	48.42
10—	103	616	11.74	60.16
5—	120	736	13.67	73.83
0—	142	878	16.17	100.00
N	878		100.00	
M	23.9			

PERCENTILE RANK DISPARITY BETWEEN PAIRED TESTS IN GEOGRAPHY

The percentile rank disparity between all tests (semester and shorter unit) in geography is shown in Table XXIV.

Groups of pupils in the fifth, sixth, and seventh grades took paired tests in geography. Table XXIV shows that the mean amount of disparity for all geography tests is 23.3. This mean is based upon the results from twenty-seven test groups.

PERCENTILE RANK DISPARITY BETWEEN PAIRED TESTS IN HISTORY

A comparison between the disparity found in history with that found in geography is of some interest. The facts for history appear in Table XXV.

Paired history tests were administered in the fifth, sixth, and seventh grades, and in high school. The mean amount of disparity for all history tests is approximately 22 percentile points (Table XXV). The lowest mean is 20.0 and the highest is 23.7. The corresponding range for geography is 21.5 to 25.4. The summary mean for geography is 23.3. The difference between geography and history tests with respect to range and mean of percentile rank differences is so small as to warrant no further comment.

TABLE XXIII. SUMMARY OF MEANS—PERCENTILE POINT DISPARITY FOR ALL TESTS WHICH COVER A SEMESTER OF WORK

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography	7	25.4	36.2	10
History	7	23.7	35.2	13
History	SHS	22.7	29.0	2
Mean of means		23.9	33.5	

TABLE XXIV. ALL GEOGRAPHY TESTS—SUMMARY OF MEAN PERCENTILE POINT DISPARITY

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography	5	21.5	61.2	10
Geography	6	22.9	64.0	7
Geography	7	25.4	36.2	10
Mean of means		23.3	53.8	

FINDINGS FROM ALL TESTS

Sixty-eight groups of pupils (3,133 individuals) were given paired tests in connection with this study of percentile rank disparity. The tests were distributed as follows: geography, twenty-seven pairs; history, forty pairs; and health, one pair. A distribution of the results from all tests is given in Table XXVI.

An analysis of the cumulative columns of Table XXVI gives the most revealing picture of the findings. Three hundred and nineteen pupils, or about 10 per cent of the group, varied 50 percentile points or more. Approximately 10 per cent in each of the distributions presented already have consistently varied 50 or more percentile points. Five hundred and sixty-two, or 17.9 per cent, showed a disparity ranging from 40 percentile points upward. Finally, one

TABLE XXV. ALL HISTORY TESTS—SUMMARY OF MEAN PERCENTILE POINT DISPARITY

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
History	5	23.0	47.0	8
History	6	21.6	49.1	9
History	7	23.7	35.2	13
History	SHS	20.0	33.9	10
Mean of means		22.1	41.3	

TABLE XXVI. ALL TESTS—SUMMARY DISTRIBUTION OF PERCENTILE POINT DISPARITY

Percentile Points	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
95—	1	1	0.03	0.03
90—	3	4	0.10	0.13
85—	5	9	0.16	0.29
80—	8	17	0.26	0.55
75—	13	30	0.41	0.96
70—	23	53	0.73	1.69
65—	43	96	1.37	3.06
60—	65	161	2.07	5.13
55—	55	216	1.76	6.89
50—	103	319	3.29	10.18
45—	101	420	3.22	13.40
40—	142	562	4.53	17.93
35—	165	727	5.27	23.20
30—	204	931	6.51	29.71
25—	230	1161	7.34	37.06
20—	301	1462	9.61	46.66
15—	351	1813	11.20	57.86
10—	352	2165	11.24	69.10
5—	489	2654	15.61	84.71
0—	479	3133	15.29	100.00
N	3133		100.00	
M	23.0			

should note that 1,161 pupils, distinctly more than one-third (37 per cent) of the group, showed a minimum disparity of 25 percentile points, one-fourth of the possible disparity.

A summary of the means for all the grade-subject groups is given in Table XXVII.

The summary mean is 23.0. This mean is based upon sixty-eight test groups. The most convincing characteristic of Table XXVII is

TABLE XXVII. ALL TESTS—SUMMARY OF MEANS OF PERCENTILE POINT DISPARITY

Subject	Grade	Mean Percentile Point Disparity	Mean N	Number of Test Groups
Geography	5	21.5	61.2	10
Geography	6	22.9	64.0	7
Geography	7	25.4	36.2	10
History	5	23.0	47.0	8
History	6	21.6	49.1	9
History	7	23.7	35.2	13
History	SHS	20.5	33.9	10
Health	6	25.8	100.0	1
Mean of means		23.0	53.3	

the consistency of the means in respect to magnitude. The means cluster very closely around the central tendency, 23.0. Such consistency seems to warrant the conclusion that the amount of disparity reported in this investigation is probably very near to that which one might find generally when objective tests are used in regular testing procedures.

CHAPTER VI

Teacher-Mark Analysis of Informal Objective Test Results

In this chapter the data which resulted from administering paired informal objective tests are analyzed in terms of teacher-mark[1] disparity. A teacher mark for a given pupil on a test is simply the letter grade to which that pupil's score on the test is equivalent. Each pupil who took paired tests had two teacher marks or letter grades. To show the extent of disparity or difference between these two sets of marks for the subjects and grade levels investigated is the purpose of this chapter.

The teacher-mark analysis is subject to two important limitations. In the first place, the teacher marks were based upon only one end of the distribution of scores. For example, the system most frequently used by those teachers who assisted with this study was based upon percentage figures distributed in the following manner: 90-100, *A;* 80-89, *B;* 70-79, *C;* below 70, *F* (failing). Thus a score of 68 on one test and a score of 34 on a second test were both equivalent to the mark *F*,[2] and in the teacher-mark analysis the two marks were considered as showing no disparity. This is obviously misleading; the effect is to decrease the actual amount of disparity. Hence the disparity reported is in reality too small.

In the second place, the same type of criticism applies when both tests were unusually easy, and, consequently, the distribution was skewed toward the upper end of the scale. Either of the conditions mentioned would tend to lower spuriously the manifest teacher-mark disparity.

TESTS WHICH COVER IDENTICAL UNITS OF TEXT MATTER

Fifth grade geography. The teacher-mark disparity between paired tests in fifth grade geography shown in Table XXVIII will serve as an illustration of the teacher-mark analysis.

The term "mark interval" which appears as the heading of column 1 in Table XXVIII is used to describe the distance from one

[1] See p. 36 for an example of the original data.

[2] In many cases the tests used in this study were so difficult (in terms of errors) that an unusually large number of the scores gave a percentage equivalent below 70. It should be clear that this fact does not affect the correlation of raw scores and the percentile rank analyses.

TABLE XXVIII. FIFTH GRADE GEOGRAPHY—DISPARITY IN TERMS OF TEACHERS' MARKS BETWEEN INFORMAL OBJECTIVE TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Mark Interval	Frequency All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3	58	58	9.5	9.5
2	109	167	17.8	27.3
1	213	380	34.8	62.1
0	232	612	37.9	100.0
N	612		100.0	
M	0.98			

letter mark to another in the scale. *A* is *one* mark interval from *B*, *two* mark intervals from *C*, and *three* mark intervals from *F*. (Only four marks were used.) All identical marks are said to show *zero* mark interval disparity. The greatest possible mark interval difference between the teacher marks from two tests is *three*.

Table XXVIII indicates that fifty-eight, or 9.5 per cent, of the pupils who took paired tests in fifth grade geography showed a disparity of three mark intervals—that is, these pupils made *F's* on one test and *A's* on another. One hundred and nine pupils, or 17.8 per cent of the group, showed a disparity of two mark intervals. An examination of the cumulative columns is illuminating. One hundred and sixty-seven, or 27 per cent, received marks two or more intervals apart. Further, 380 pupils, 62 per cent of the group, received marks which varied one or more mark intervals. The mean mark-interval disparity for fifth grade geography is .98 mark intervals or approximately *one*.

Other identical unit tests. Facts[3] similar in nature to those just presented were secured for sixth grade geography, fifth and sixth grade history, high-school history, and sixth grade health. The consistency of the percentages of the pupils who varied one or more mark intervals is of interest. It will be recalled that this percentage indicates the per cent of pupils whose marks varied one or more mark intervals. Sixty-two per cent of the fifth grade geography pupils varied to this extent. The corresponding figures for the remaining subject-grade groups are as follows: sixth grade geography, 62.5; fifth grade history, 72.9; sixth grade history, 65.4; high-school history, 50.8; sixth grade health, 82.0.

Summary for all identical text matter tests. Teacher marks were

[3] Complete data for each subject is not presented due to space limitations. Anyone desiring such facts should correspond with the author.

secured on paired objective tests for forty test groups (2,134 pupils). Summaries of the data found are shown in Tables XXIX and XXX. Table XXIX is a summary distribution of mark-interval disparity based upon the forty test-group distributions.

TABLE XXIX. SUMMARY DISTRIBUTION—TEACHER-MARK DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Mark Interval	Frequency–All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3	136	136	6.4	6.4
2	446	582	20.9	27.3
1	796	1378	35.4	62.7
0	756	2134	37.3	100.0
N	2134		100.0	
M	0.96			

This table reveals that 136, or 6 per cent of 2,134 individuals, varied three mark intervals on paired tests. Four hundred and forty-six, or approximately one-fifth of the group, showed a teacher-mark disparity of two mark intervals. Thirty-five per cent of the individuals received marks one mark interval apart. The cumulative percentage column shows that somewhat more than one-fourth of the pupils were given marks which differed two or more mark intervals. Further, almost two-thirds of the group (62.7 per cent) showed a minimum teacher-mark disparity of one or more mark intervals.

A summary in terms of mean mark disparity for subject-grade groups is presented in Table XXX.

TABLE XXX. SUMMARY OF MEANS—TEACHER-MARK DISPARITY FOR ALL TESTS WHICH COVER IDENTICAL TEXT MATERIAL

Subject	Grade	Mean Teacher-Mark Disparity	Mean N	Number of Test Groups
Geography	5	0.98	61.2	10
Geography	6	0.87	64.0	7
History	5	1.03	47.0	8
History	6	1.02	51.2	8
History	SHS	0.76	32.8	6
Health	6	1.03	100.0	1
Mean of means		0.96	59.4	

The most significant feature of the data in Table XXX is the consistency in size of the means. The lowest mean is .76 and the

highest is 1.08. If the somewhat special case[4] of high-school history is disregarded for the moment, the lowest mean becomes 0.87. The summary mean is 0.96 or approximately 1.0. One should recall here that because of the limitations of this technique 3.0 is the greatest possible disparity and that, therefore, the mean amount of disparity found, *one interval,* is one-third of the total amount possible.

TESTS WHICH COVER A SEMESTER OF WORK

Seventh grade geography, seventh grade history, and high-school history. Teacher marks were computed for groups of pupils in seventh grade geography, in seventh grade history and in high-school history. The teacher-mark disparity for each of these subjects was determined. The distributions are very similar in nature to those already presented, except that the amount of disparity tends to be somewhat greater for semester tests. Again attention is called to the marked consistency in the shape of the distributions. In seventh grade geography 72 per cent varied *one or more* mark intervals, in seventh grade history 69 per cent, and in high-school history 70 per cent.

Summary for all semester tests. Twenty-five groups of pupils (878 individuals) took paired semester tests. Summaries of the teacher-mark disparity revealed are given in Tables XXXI and XXXII.

TABLE XXXI. SUMMARY DISTRIBUTION—TEACHER-MARK DISPARITY FOR ALL TESTS WHICH COVER A SEMESTER OF WORK

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3	97	97	11.0	11.0
2	241	338	27.4	38.4
1	283	621	32.3	70.7
0	257	878	29.3	100.0
N	878		100.0	
M	1.28			

A comparison of Table XXXI with the corresponding table (Table XXIX) for identical unit tests indicates that on the whole the disparity is somewhat greater in the case of semester tests, but the difference is hardly pronounced enough to establish a trend. Eleven per cent of the pupils varied three mark intervals on semester

[4] See pp. 40-41 for further discussion. Also, the fact that an unusually large number of pupils made below 70 on both pairs of tests tends to explain this figure.

tests, as compared with 6.4 per cent on the shorter identical unit tests. Thirty-eight per cent showed a disparity of two or more intervals on semester tests, whereas the corresponding figure for identical unit tests was 27. Finally, 70.7 per cent of the children varied a minimum of *one* mark interval on semester tests and 62 per cent varied the same amount on the other type of test.

A summary of the mean teacher-mark disparity for semester tests is shown in Table XXXII.

TABLE XXXII. SUMMARY OF MEANS—TEACHER-MARK DISPARITY OF ALL TESTS WHICH COVER A SEMESTER OF WORK

Subject	Grade	Mean Teacher-Make Disparity	Mean N	Number Test Groups
Geography	7	1.15	36.2	10
History	7	1.20	35.2	13
History	SHS	1.49	29.0	2
Mean of means		1.28	33.5	

The summary mean is 1.28 mark intervals. This figure may be, to a small degree, spuriously high due to the fact that the mean for the high-school tests is based upon a relatively small number of cases and, therefore, may not be representative. However, one should recall that the semester high-school tests were the longest and perhaps the most carefully constructed tests used in the study. Nevertheless, the facts at hand show that teacher marks on paired objective tests varied about one and one-fourth mark intervals when semester tests were used.

GEOGRAPHY AND HISTORY TESTS

All geography tests. The teacher-mark disparity for all geography tests is shown in Table XXXIII.

TABLE XXXIII. ALL GEOGRAPHY TESTS—SUMMARY DISTRIBUTION OF TEACHER-MARK DISPARITY

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3	90	90	6.3	6.3
2	296	386	20.8	27.1
1	536	922	37.7	64.8
0	500	1422	35.2	100.0
N	1422		100.0	
M	1.0			

Note the percentage column. Approximately 6 per cent of the pupils made *F* on one test and *A* on another; about one-fifth (20.8 per cent) showed a disparity of two marks; and 37 per cent varied one mark interval. Almost two-thirds of the group received marks one or more mark intervals apart. The mean for geography tests is 1.0 mark interval.

All history tests. Similar facts for all history tests are given in Table XXXIV. The distributions for history and geography are very similar. For example, in geography 64.8 per cent of the pupils varied a minimum of one mark interval; the corresponding figure for history is 66.8. The mean for history tests is somewhat higher, due to a slightly greater proportion of cases in the two- and three-interval arrays.

TABLE XXXIV. ALL HISTORY TESTS—SUMMARY DISTRIBUTION OF TEACHER-MARK DISPARITY

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3	135	135	9.1	9.1
2	359	494	24.1	33.2
1	501	995	33.6	66.8
0	495	1490	33.2	100.0
N	1490		100.0	
M	1.11			

SUMMARY FOR ALL TESTS

The summary distribution given in Table XXXV includes all of the pupils for whom teacher marks were available. The data are based upon 3,012 cases (sixty-five test groups).

TABLE XXXV. ALL TESTS—SUMMARY DISTRIBUTION OF TEACHER-MARK DISPARITY

Mark Interval	Frequency—All Tests	Cumulative Frequency	Percentage Distribution All Tests	Cumulative Percentage Distribution
3	233	233	7.7	7.7
2	687	920	22.8	30.5
1	1079	1999	35.8	66.3
0	1013	3012	33.7	100.0
N	3012		100.0	
M	1.04			

Table XXXV shows that 233 pupils, or 7.7 per cent, varied from one test to another to the extent of three mark intervals. About

one-fifth of the pupils showed a disparity of two mark intervals; approximately one-third of the pupils one mark interval; and one-third received identical marks. The cumulative percentage column in Table XXXV shows the facts very clearly. Almost one-third (30.5 per cent) of the pupils varied two or more mark intervals, and two-thirds of the group (66.3 per cent) varied a minimum of one mark interval.

A summary of the mean teacher-mark disparity found appears in Table XXXVI.

TABLE XXXVI. ALL TESTS—SUMMARY OF MEANS OF TEACHER-MARK DISPARITY

Subject	Grade	Mean Teacher-Mark Disparity	Mean N	Number of Test Groups
Geography	5	0.98	61.2	10
Geography	6	0.87	64.0	7
Geography	7	1.15	36.2	10
History	5	1.08	47.0	8
History	6	1.02	51.2	8
History	7	1.20	35.2	13
History	SHS	1.12	30.9	8
Health	6	1.03	100.0	1
Mean of means		1.06	53.2	

Table XXXVI gives the mean teacher-mark disparity for subject-grade groups. The highest of the means is 1.20 and the lowest is 0.87, a range of 0.33. As a whole, however, these means center very closely about the central tendency, 1.06. Note that six of the means fall between 0.98 and 1.15.

Concluding statement. The facts presented in Tables XXXV and XXXVI warrant the following conclusions: when teacher marks are assigned to pupils on the basis of two objective tests constructed by competent teachers to measure knowledge of the same subject matter the two marks will vary an average of approximately one mark interval.

CHAPTER VII

DESCRIPTION OF PROCEDURES USED IN STUDY OF STANDARDIZED TESTS

The purpose of this part of the investigation was the same as that of the teacher-made test study, namely, to determine the extent of disparity between tests which were designed to measure the same abilities. In this chapter the procedures which pertain to the study of standardized commercial tests are described.

Selection of tests. In order to ascertain the disparity between comparable commercial tests, it was necessary to administer two or more tests to the same children. Although the results are analyzed in terms of specific subject tests and not in terms of batteries, in order to simplify administration the sub-tests of certain batteries were chosen for investigation.

Three of the better known and more complete batteries were chosen for study, namely, The New Stanford, Advanced Examination, Form W; The Metropolitan Achievement Tests, Intermediate Battery—Complete, Form A; and The Public School Achievement Test, Batteries A, B, and C, Form 3. In order to facilitate description the three batteries are given the following designation: Metropolitan, *a;* Public School, *b;* New Stanford, *c.* Thus, "Reading *a*" is the reading test from the Metropolitan battery.

Table XXXVII shows the number and type of sub-tests in each battery.

As shown in Table XXXVII, three comparisons were possible in each of seven subjects, and one comparison could be made in each of the remaining three subjects. There was a possible total then of twenty-four subject-test comparisons.

Administration of tests. The tests were administered in the sixth grades of seven public schools in Durham, North Carolina. Four hundred and sixty pupils took the three tests.

The three batteries were given as a final examination on a semester of school work. They were administered during the regular examination period, and were considered by the pupils as a measure of their progress. This condition provided a relatively strong motiva-

TABLE XXXVII. SUB-TEST COMPOSITION OF METROPOLITAN, NEW STANFORD, AND PUBLIC SCHOOL ACHIEVEMENT TESTS

Subject	Battery			Number of Comparisons Possible
	a	b	c	
Reading	*	*	*	3
Vocabulary	*	..	*	1
Arithmetic: Reasoning	*	*	*	3
Computation	*	*	*	3
English usage	*	*	*	3
Literature	*	..	*	1
History	*	*	*	3
Geography	*	*	*	3
Spelling	*	*	*	3
Health	..	*	*	1
Total	9	8	10	24

*Indicates that there is a sub-test.

tion. Three days were set aside by the school system during which the sixth grade children were freed from all other routine school activities. The pupils were permitted to go home at the end of the testing day. The children manifested a keen interest in the tests throughout the testing period. In fact, they seemed, in the main to enjoy the taking of these tests, and in many cases overtly expressed regret that it was necessary to return to the school routine.

In order to minimize fatigue and practice effects, three administrative precautions were taken. First, the schools were paired according to the general type of children in attendance. A school having a large number of slow or retarded children in attendance was paired with a school having a relatively large number of "bright" or accelerated pupils to make an "administrative group." This guaranteed that each administrative group would consist of representative children. In the discussion which follows the administrative groups have the following designation: Group I (school 1 and school 7); Group II (school 5, school 2, and school 3); and Group III (school 4 and school 6).

As a second precaution, a method of rotation was used in administering the tests. The pertinent facts concerning this method are presented in Table XXXVIII.

Hence, if any position on the program was favorable the effect should have been neutralized when the three groups were considered as a whole.

The third precaution pertained to sequence of sub-tests. As has

TABLE XXXVIII. METHOD OF ROTATION USED IN ADMINISTERING STANDARDIZED TESTS

Administrative Group	Date Tests Were Administered		
	January 14	January 15	January 16
Group I	Metropolitan	New Stanford	Public School
Group II	New Stanford	Public School	Metropolitan
Group III	Public School	Metropolitan	New Stanford

been stated the sub-tests were considered as separate subject[1] tests. Since the results from these tests were to be compared, it was necessary that the individual sub-tests in a given subject field be given in a manner as nearly as possible identical. In order to effect this condition, the sub-tests in each battery were given in the same sequence, and consequently the tests to be compared came at approximately the same time of day. For example, in every case the Arithmetic Computation test was given as the third test, appearing at the beginning of the second sitting. Thus, if there was any advantage or disadvantage in this particular place in the day's program, the effect should have been the same for each of the three tests to be compared.[2]

The children were given brief rest periods between sub-tests and were allowed a play period in the open air between the four major sittings.

The tests were administered by persons who were trained and experienced in the giving of standardized tests. Special emphasis was placed upon an exact adherence to the manual of instructions. In order to insure proper administration, a special conference was held with each person who assisted with the administration of the tests. In this conference the purpose of the study was carefully explained, and the manuals and tests were examined in detail. The point that the same person administered all the tests to a given

[1] Many of the sub-tests of the batteries used in this investigation are sold as separate tests.

[2] One should keep in mind that the purpose of this study was to *compare* the results from standardized commercial tests. Whether or not the conditions were most favorable for excellence of performance is not of fundamental significance for this study provided the conditions were the same for each of the tests to be compared. The criticism may be made that the testing program was too strenuous, but neither facts nor observation seems to support this contention. For example, the facts seem to indicate that the relationships between the scores on the sub-tests of different batteries were relatively unchanged regardless of the sequence in which the batteries were given. However, this point does not affect the validity, for the purposes of this study, of the rotation method used.

group of pupils is worthy of note, for this offset any advantage or disadvantage that might have accrued from the personality of a particular individual tester.

Scoring the tests. The usual precautions were taken to guarantee accuracy in the scoring of the tests. The factors given special consideration are listed here. (1) A central scoring place was used in order to insure complete uniformity. (2) All persons who assisted with the scoring had experience in scoring standardized tests. (3) The whole scoring procedure was carried on under the constant supervision of the investigator. (4) All scoring was done in strict accordance with keys and manuals. (5) In order that groups of papers revealing material errors might be re-scored, samples of each group[8] of papers were carefully examined.

All raw scores were transmuted into grade-equivalent norms which are furnished by the publishers of the tests. In every case this transmutation was performed in strict accordance with manual directions. These grade equivalents are calculated in terms of one-tenths of a school year. Thus a grade equivalent of 6.4, although often interpreted as representing achievement expected at the end of the fourth month of the sixth grade, may be more technically considered as representing six and four-tenths years of school work. The procedures relative to the calculation of grade equivalents are uniform for all the tests.

Nature and treatment of results. When the grade equivalents for any given tests (for example, the three reading tests) had been determined, the three measures of each child's ability were expressed in comparable terms. A given child, *x*, had grade equivalents as follows in reading:

Ra 6.0
Rb 8.5
Rc 7.4

The problem of Part B of the study was to determine the extent of variation or disparity between the performance of pupils on comparable standardized commercial tests. Two methods of analysis were used to show the extent of disparity. First, the scores from paired tests were correlated (Chapter VIII); second, the difference or disparity in test results was calculated in terms of months or one-tenths of a school year (Chapter IX).

[8] Those papers scored by a given scorer, or those sub-tests scored by a given individual.

CHAPTER VIII

Correlation Analysis of Standardized Commercial Test Results

A correlation coefficient may be interpreted in several ways, but perhaps the simplest and most meaningful interpretation for the present purpose is that which states that the square of the coefficient indicates the percentage of factors common to the correlated variables.[1] Obviously, a correlation of 1.00 would mean that 100 per cent of the factors producing the scores were common. By the same method, it is clear that a coefficient of .50 (.25 when squared) would indicate the presence of 25 per cent of overlapping or identical factors.

A correlation table will further clarify the meaning of the coefficients.[2] Table XXXIX shows the distribution of cases when r is .68.

The cases which appear in a given interval on one test are distributed widely on the other test. As an example, note the distribution of the sixty cases in the interval 6.0-6.3 on Test *a* (Table XXXIX). These cases are distributed on Test *b* in intervals ranging from 4.8-5.1 to 9.6-9.9. Only eight of the sixty cases appear in the 6.0-6.3 interval on Test *b*. Similar analysis of other arrays enables one to secure a clearer understanding of the individual variation involved in a correlation of .68.

Correlations for all subject tests. If two standardized commercial tests constructed by experts to measure the same ability are given to the same children under comparable conditions, what will be the relationship between the two sets of scores secured? The answer to this question for the sub-tests of the three standardized tests is given in Table XL.

One should note that the subject tests are arranged alphabetically in Table XL. The median of the coefficients is .68. The corresponding figure for teacher-made tests is .54. Thus, although standardized commercial tests show somewhat less disparity than teacher-made

[1] This is a widely used interpretation. See Henry E. Garrett, *Statistics in Psychology and Education* (New York: Longmans, Green, and Company, 1926), pp. 291-298.

[2] For a valuable discussion of this type of analysis see Frank Sandon, "The Necessary Imperfections of An Examination," *The British Journal of Educational Psychology,* V (June, 1935), 191-192.

TABLE XXXIX. CORRELATION TABLE—READING TESTS *a* AND *b*

Test *a*	Test *b* 4.0–4.3	4.4–4.7	4.8–5.1	5.2–5.5	5.6–5.9	6.0–6.3	6.4–6.7	6.8–7.1	7.2–7.5	7.6–7.9	8.0–8.3	8.4–8.7	8.8–9.1	9.2–9.5	9.6–9.9	F
10.0–10.3	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
9.6– 9.9	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
9.2– 9.5	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
8.8– 9.1	....	....	....	....	....	....	....	....	....	....	....	....	....	....	2	2
8.4– 8.7	....	....	....	....	....	....	....	....	....	....	1	....	1	....	5	7
8.0– 8.3	....	....	....	....	....	....	....	....	1	....	3	3	4	7	10	28
7.6– 7.9	....	....	....	....	....	1	....	....	9	1	6	5	1	2	6	31
7.2– 7.5	....	....	....		1	1	2	6	10	4	17	16	5	2	4	68
6.8– 7.1	....	....	....	1	1	5	5	1	14	4	10	6	1	2	....	50
6.4– 6.7	....	....	....	3	2	6	7	9	12	4	6	2	1	1	....	53
6.0– 6.3	....	....	1	4	5	8	12	5	7	2	9	3	1	1	2	60
5.6– 5.9	....	....	....	8	7	8	8	5	9	3	3	3	....	....	1	55
5.2– 5.5	1	1	3	11	6	5	4	2	9	2	3	1	....	....	....	48
4.8– 5.1	2	4	4	8	4	8	2	3	7	....	....	2	....	1	....	45
4.4– 4.7	3	....	2	4	....	2	....	1	....	....	....	....	....	1	....	13
F	6	5	10	39	26	44	40	32	78	20	58	41	14	17	30	....

N 460
r .675

objective tests, it is clear that a large degree of variability remains in the reputedly more refined standardized tests.

As indicated by Table XL, the geography tests *b* and *c* show the highest correlation. The lowest coefficient is that for health tests (.496). The relatively high correlation between Tests *b* and *c* in geography is misleading in a sense, for its unusual size was due to a relatively small number of cases—about 25—who made the highest possible score on both tests. If these cases are omitted, the correlation becomes approximately the same as that found for the other geography tests, namely, about .65.

The column entitled r^2 in Table XL may be considered as one index to the disparity in results from the tests involved. If the correlation is .70 or below, less than 50 per cent of the factors which produced the scores may be said to be identical. Fourteen of the twenty-four correlations come in this group.

Closely related studies. About the time the investigation just described was completed, Foran and Loyes[3] reported a very similar

[3] T. G. Foran and Sister M. Edmund Loyes, "The Relative Difficulty of Three Achievement Examinations," *The Journal of Educational Psychology*, XXVI (March, 1935), 218-222.

TABLE XL. CORRELATIONS BETWEEN STANDARDIZED ACHIEVEMENT TESTS WHICH WERE DESIGNED TO MEASURE THE SAME ABILITIES
N 460

Subjects	Test Correlated*	r	r^2 (Per Cent of Common Factors)
Arithmetic computation	*a* and *b*	.749	.56
Arithmetic computation	*b* and *c*	.696	.48
Arithmetic computation	*a* and *c*	.708	.50
Arithmetic reasoning	*a* and *b*	.708	.50
Arithmetic reasoning	*b* and *c*	.713	.51
Arithmetic reasoning	*a* and *c*	.635	.40
Geography	*a* and *b*	.664	.44
Geography	*b* and *c*	.906	.82
Geography	*a* and *c*	.576	.33
Health	*b* and *c*	.496	.25
History	*a* and *b*	.590	.35
History	*b* and *c*	.643	.41
History	*a* and *c*	.565	.32
Language usage	*a* and *b*	.647	.42
Language usage	*b* and *c*	.649	.42
Language usage	*a* and *c*	.645	.42
Literature	*a* and *c*	.588	.34
Reading	*a* and *b*	.675	.46
Reading	*b* and *c*	.705	.50
Reading	*a* and *c*	.730	.53
Word meaning	*a* and *c*	.790	.62
Spelling	*a* and *b*	.863	.74
Spelling	*b* and *c*	.848	.72
Spelling	*a* and *c*	.844	.71
Median	.....	.68	...

*Test *a* Metropolitan, *b* Public School, *c* New Stanford.

study. The tests used were: (1) The New Stanford, Advanced Examination, Form V; (2) The Modern School Achievement, Test I; and (3) The Unit Attainment Scale, Form A, Division 2. The results from the three tests were correlated. The coefficients found are presented in Table XLI.

The median of the twenty-three coefficients in Table XLI is .60. The median coefficient for the Durham study was .68. Foran's and Loyes's facts further illuminate the problem of this investigation. If the data from the two studies are combined, some tentative generalizations concerning the relationship between standardized tests in certain subjects at the grammar grade level may be ventured. The mean correlation for the various subject tests (based upon coefficients from both studies) is shown in Table XLII.

The mean for all subject tests is .633. The median of the forty-seven coefficients is .647. Note the mean for specific subjects as shown in Table XLII. Arithmetic computation, for example, shows

TABLE XLI. CORRELATIONS FROM FORAN'S AND LOYES'S STUDY OF THREE STANDARDIZED TESTS*

Subject	Tests Correlated**	r
Arithmetic computation	Unit and M.S.	.564
	Unit and N.S.	.605
	M.S. and N.S.	.709
Arithmetic reasoning	Unit and M.S.	.603
	Unit and N.S.	.592
	M.S. and N.S.	.611
Geography	Unit and M.S.	.661
	Unit and N.S.	.678
	M.S. and N.S.	.898
Health	M.S. and N.S.	.440
History	Unit and M.S.	.514
	Unit and N.S.	.572
	M.S. and N.S.	.545
Language usage	Unit and M.S.	.428
	Unit and N.S.	.239
	M.S. and N.S.	.455
Literature	Unit and N.S.	.412
Reading—Paragraph M.	Unit and M.S.	.543
	Unit and N.S.	.636
	M.S. and N.S.	.725
Spelling	Unit and M.S.	.670
	Unit and N.S.	.685
	M.S. and N.S.	.745
Median	.....	.603

*Data adapted from T. G. Foran and Sister M. Edmund Loyes, *op. cit.*
**Unit—Unit Attainment Scale, Form A, Division 2.
M.S.—Modern School Achievement, Test I.
N.S.—New Stanford, Advanced Examination, Form V.

a mean correlation of .672, based upon the correlation of six pairs of standardized tests designed to measure the same ability. The mean for other subjects may be read from Table XLII.

TABLE XLII. MEAN CORRELATIONS FOR STANDARDIZED TESTS IN VARIOUS SUBJECTS BASED UPON DATA FROM FORAN AND LOYES AND UPON DATA FROM THE PRESENT STUDY

Subject	Mean r	Range of r's	Number r's in Subject Goup
Arithmetic computation	.672	.564 — .749	6
Arithmetic reasoning	.644	.592 — .713	6
Geography	.730	.576 — .906	6
Health	.468	.440 — .496	2
History	.571	.514 — .643	6
Language usage	.510	.239 — .649	6
Literature	.500	.412 — .588	2
Reading	.669	.543 — .730	6
Spelling	.776	.670 — .863	6
Word meaning	.790	.....	1
N	....	.....	47
Mean of means	.633	.....	..

Ruch and others[4] correlated the scores on ten American History tests. The mean correlation between a given history test and nine other history tests (for each of the ten tests studied) is shown in Table XLIII.

TABLE XLIII. MEAN CORRELATIONS BETWEEN AMERICAN HISTORY TESTS (STANDARDIZED) BASED UPON DATA FROM RUCH AND OTHERS

Test Form	Mean r With Other Nine Tests
1. Gregory A	.66
2. Gregory B	.63
3. Barr A	.54
4. Barr B	.61
5. Pressey and Richards	.67
6. Van Wagenen S-3	.60
7. Kepner A	.67
8. Kepner B	.60
9. Van Wagenen History Reasoning A	.48
10. Van Wagenen History Reasoning B	.48
Mean of means	.59
Median of means	.60

*Data adapted from Henry L. Smith and Wendell W. Wright, *Tests and Measurements* (New York: Silver, Burdett and Company, 1928), p. 242.

The median of these means is .60 and the mean is .59. These figures are very similar in magnitude to corresponding figures already presented.

Concluding statement. Data available warrant the following conclusion: When two standardized achievement tests constructed and standardized by experts for the purpose of measuring the same abilities are given to pupils and the scores correlated, the median correlation tends to be approximately .65. This means that about 42 per cent of the factors in the testing situation are identical; or stated in other terms, that about 58 per cent of the factors which produce the scores are different.

[4] G. M. Ruch, M. H. De Graff, W. E. Gordon . . . (and others), *Objective Examination Methods in the Social Studies* (New York: Scott, Foresman and Company, 1926).

CHAPTER IX

Analysis of Standardized Test Findings in Terms of Grade-Equivalent Disparity

A coefficient of correlation does not reveal the extent and nature of *individual* variations between the scores on two tests. Therefore, in order to show more clearly the disparity between the standardized objective tests used in this study, the findings are analyzed in this chapter in terms of grade-equivalent scores—commonly expressed in years and months.

A brief illustration will serve to clarify the basic data on which the tables presented in this chapter are based. The following are the grade-equivalent scores of three pupils on the three standardized tests in reading (paragraph meaning).

Pupil	Test *a*	Test *b*	Test *c*
1	7.2	9.0	8.7
2	7.4	8.2	7.0
3	6.9	6.0	7.6

The disparity between Tests *a* and *b* was calculated in the following manner. Test *b* was considered as a base[1] and the disparity between the two tests was the number of months that the Test *a* scores varied *from* the Test *b* scores. Whether the variation was above or below the base score was not taken into consideration. For pupil 1 (in the illustration) the variation (*a* from *b*) was eighteen months (9.0-7.2). The tables which follow contain distributions of these individual variations for all subject tests given.

Grade-equivalent disparity—all subjects

Arithmetic computation. The grade-equivalent disparity for the three arithmetic computation tests is shown in Table XLIV.

Table XLIV indicates that the mean amount of disparity between Tests *a* and *b* in arithmetic computation is 8.4 months. Approximately 35 per cent of the pupils had grade-equivalent scores on Test *a* which were one school grade (ten months) or more from their grade position as determined by Test *b*. The mean disparity between Tests *a* and *c* is almost two school grades (18.5 months). About 80

[1] In the case of a given comparison, the test to be used as a base was chosen arbitrarily. The choice in no way affects the findings, for the variation is simply the distance in months from one score to another.

TABLE XLIV. ARITHMETIC COMPUTATION—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST *a* FROM TEST *b*		TEST *a* FROM TEST *c*		TEST *b* FROM TEST *c*	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
42–44	..	....	1	0.22	..	....
39–41	..	....	2	0.65	1	0.22
36–38	..	....	8	2.39	2	0.65
33–35	..	....	11	4.78	3	1.30
30–32	1	0.22	25	10.22	6	2.61
27–29	3	0.87	30	16.74	20	6.96
24–26	3	1.52	52	28.04	17	10.66
21–23	5	2.61	57	40.43	26	16.31
18–20	20	6.96	58	53.04	45	26.09
15–17	32	13.92	62	66.52	41	35.00
12–14	42	23.05	45	76.30	58	47.61
9–11	83	41.09	35	83.91	68	62.39
6–8	80	58.48	25	89.35	62	75.87
3–5	104	81.09	31	96.09	69	90.87
0–2	87	100.00	18	100.00	42	100.00
Total	460		460		460	
Mean	8.4		18.5		12.8	

per cent of the pupils varied one grade or more. Note further that almost 10 per cent of the pupils were thirty months or more apart on the two tests. This means that these pupils had grade-equivalent scores on two standardized tests such as the following: Test *a*, 5.0, Test *b*, 8.0; Test *a*, 7.0, Test *b*, 4.0. The full meaning of this variation is clearer if one recalls that the lowest *possible* grade-equivalent is 0 and the highest in this case is 10.0; that these pupils were all in the sixth grade; and that both of the tests were designed to measure arithmetic computation ability.

The mean variation of Test *b* from Test *c* was approximately one- and three-tenths school grades (12.8 months). Nearly 20 per cent of the groups showed a minimum disparity of *two* grades, and more than 50 per cent a minimum of one school grade.

The mean of the three arithmetic computation means is 13.2 months or about one and three-tenths school grades. On the average, then, these three standardized tests showed a disagreement more than one school grade in respect to the arithmetic computation achievement of the 460 pupils.

Other subject tests. The data which pertain to grade-equivalent disparity or variation for standardized tests in arithmetic reasoning, geography, health, history, language usage, literature, reading, word meaning, and spelling are presented in Tables XLV* to LIII*, in-

clusive. The facts in these tables are similar in kind to those given and analyzed for arithmetic computation. The mean amount, the range, and the distribution of the disparity found in the case of particular subjects may be read from the respective tables.

Summary of grade-equivalent data. A summary of the mean grade-equivalent disparity for all subject tests is given in Table LIV.

TABLE LIV. SUMMARY OF MEAN GRADE-EQUIVALENT DISPARITY FOR ALL SUBJECT TESTS

Subject	Mean of means — Disparity in Months
Arithmetic computation	13.2
Arithmetic reasoning	7.8
Geography	11.0
Health	12.1
History	11.1
Language	12.9
Literature	10.7
Reading—Paragraph meaning	10.0
Spelling	5.7
Word meaning	7.2
Mean of means	10.2

With the exception of three subjects, the means in Table LIV are very consistent. The spelling tests clearly showed less variation than any of the other tests given. The means for arithmetic reasoning and word meaning are significantly lower than the mean of the means. It is of interest, however, to note that even in the case of spelling the mean disparity is somewhat more than one-half of a school grade. The mean disparity for seven of the ten subject fields ranges from one school grade (10 months) to one and three-tenths school grades (13.2 months).

The mean of the means, including spelling, is 10.2 months or approximately one school grade. This fact signifies that if the 460 children were classified by one of the standardized tests and then by a second or a third supposedly comparable test, on the average the individual children would be classified on the second or third test one school grade or ten months from their positions on the first test. Thus, if a pupil made a grade-equivalent score of 6.4 on the New Stanford arithmetic computation test the findings of this study indicate that his score on either of the other two arithmetic computation tests would, in general, be either 5.4 or 7.4.

Concluding statement. In terms of grade-equivalent scores the mean amount of disparity between two standardized objective tests

constructed by experts to measure the same ability, administered in comparable fashion, and scored objectively was about one school grade (ten months). The conclusion may be drawn, therefore, that elements which produce variation enter into the testing situation in such degree as to cause an average of one school grade of disparity between the results from the two measuring instruments.

The extent of the disparity found between test results may be clarified by an illustration from the physical sciences. For example, what would the amount of disparity found in the case of objective tests mean in terms of weight? If the children averaged about sixty pounds and were distributed from thirty to one hundred pounds, there would be a mean difference between the weights of individual pupils as determined by two standard weighing instruments of approximately ten pounds. That is, if a given child weighed seventy pounds on one set of scales, and the disparity were the same as that found for tests, on the average, he would weigh sixty or eighty pounds on the second set of scales.

CHAPTER X

Summary and Conclusion

SUMMARY

1. This investigation dealt with new-type or objective achievement tests which were constructed to measure the same abilities and which were administered under comparable conditions. The purpose of the study was to determine the extent of disparity between the results from these tests. Both teacher-made and standardized objective tests were studied.

2. Sixty-three informal or teacher-made objective tests were constructed by thirty-five teachers. The subjects covered were fifth, sixth, and seventh grade geography; fifth, sixth, and seventh grade and high-school history; and sixth grade health. The tests were matched in such a manner that two tests which were constructed to measure pupil acquaintance with the same body of subject matter were considered as a "pair." Paired tests were administered to sixty-eight groups of children. Forty-three groups took tests which covered relatively short units of identical text matter, and the remaining twenty-five groups took semester examinations based upon practically identical text material. Disparity was measured in three ways.

a. When the pupil scores on paired informal tests were correlated the coefficients ranged from .845 to – .212. The median of the sixty-eight coefficients was .54. The extent of relationship was slightly less for semester tests than for tests covering a shorter unit of work.

b. A study was made of disparity in terms of the difference between percentile rank positions on the two texts. The mean percentile rank disparity found was 23 percentile points. The differences ranged from 0 to 99. About one-tenth of the pupils achieved ranks on one test which were 50 or more percentile points from their ranks on a second test; approximately one-fifth of the pupils varied in rank 40 or more percentile points; and finally, one-third of the pupils showed a minimum percentile rank difference of 25 percentile points.

c. In terms of teacher marks the disparity between paired informal objective tests was found to be approximately one mark interval. The greatest amount of difference possible was three mark

intervals. Approximately 8 per cent of the pupils varied three mark intervals (a mark of *A* on one test and a mark of *F* on a second); 23 per cent varied two mark intervals; and 36 per cent varied one mark interval. Two-thirds of the pupils received marks on paired tests one or more mark intervals apart.

3. Three standardized objective tests in each of seven subjects and two such tests in each of three subjects were administered to 460 pupils in the sixth grade. The disparity between the results from comparable tests was found (a) in terms of correlation and (b) in terms of difference in grade-equivalent scores.

a. The correlation between results from standardized tests constructed to measure the same abilities ranged from .906 to .496. The median of the twenty-four coefficients was .68. As a group the spelling tests showed distinctly higher relationship than did other subject tests.

b. In terms of grade-equivalent scores the mean amount of disparity between two standardized tests constructed by experts to measure the same abilities, administered in comparable fashion and scored objectively was found to be about one school grade (ten months). The range of disparity was from zero to somewhat more than six school grades (sixty-two months). From about 1 to 44 per cent (depending upon the subject) of the pupils varied two or more school grades.

CONCLUSION

The findings of this investigation permit certain tentative generalizations which bear upon the characteristics of objective or new-type tests as these tests are customarily used. The validity of these generalizations is dependent upon the degree to which the findings revealed by this study are representative. The procedures used in the present investigation and the consistency of the findings are submitted as evidence that the results of this study are reasonably reliable. The generalizations follow.

1. A test may be objective in the sense that all personal opinion is eliminated in scoring and still fail to remove important personal elements from the evaluation of pupil achievement. There are many factors (other than scoring) in the total measurement situation which cause marked disparity between the results from two or more new-type or objective tests constructed to measure the same functions.

2. Measures of pupil achievement obtained from different informal objective tests may be expected to vary to a considerable extent.

Thus if a pupil takes Teacher A's test his score, rank, and mark may be very different from what his score, rank, and mark would have been had he taken Teacher's B's test. This condition is to be expected even when the tests cover identical bodies of subject matter and are designed to measure the same achievement. The extent of disparity which, in general, may be expected has been expressed in the preceding summary as points *a, b,* and *c* under 2.

3. Pupil ratings based upon standardized test scores show marked disparity. Thus a grade-equivalent rating for a given child in a particular subject as determined by one standardized test may differ significantly from his grade-equivalent rating as determined by a second standardized test. Although the differences will vary from subject to subject, in general, the disparity or difference may be expected to be approximately the amount reported in the preceding summary (points *a* and *b* under 3).

PART III. PRACTICAL IMPLICATIONS AND THEORETICAL PROBLEMS

CHAPTER XI

Educational Implications and Problems for Further Research

EDUCATIONAL IMPLICATIONS

Objective tests (teacher-made and standardized) are widely used in the public schools. There has not been much evidence adduced to establish the extent to which such tests are reliable measuring instruments when the processes involved in the whole measurement situation are considered. In the absence of such evidence the tests have been uncritically accepted, and this practice has tended to foster error in the interpretation of test results. An acquaintance with the limitations of objective tests should enable the tester to make allowances and to increase the comprehensiveness of his measurements.

An example of the questionable manner in which standardized objective tests are frequently used will indicate the value of recognizing the limitations of such tests. The seventh grade pupils in the public schools of North Carolina for several years were given the New Stanford Achievement Test. The suggestion[1] was made that the grade equivalent 7.0 be taken as the minimum for promotion to the next grade. Now suppose that the State Department of Education had chosen the Public School Achievement Test instead of the test actually adopted. The facts revealed by this study indicate that in this case there would have been an average individual change in grade equivalence, for any given subject, of one school grade (ten months). For example, on the average, pupils who made a grade-equivalent score of 7.4 on a sub-test of the Stanford Achievement Test would have achieved a grade equivalent of 6.4 or 8.4 on the corresponding Public School Achievement Test. (It should be remembered in this connection that both of these tests are widely used, that both were devised and standardized by experts, that both were constructed

[1] The grade equivalent suggested was tentative and was derived from local norms (state test scores), but the point here made would be the same regardless of the specific grade-equivalent score chosen for graduation. The issue is: How dependable is a given grade equivalent as determined by a particular test?

to measure the same abilities, and that grade-equivalent standards were established in the same manner.) In problems of promotion and of grade placement, the extent and consequences of the disparity here described are obvious. It is important then that testers have the knowledge that an individual grade-equivalent score is affected in considerable degree by subjective factors entering into the construction and use of the particular test on which the score is based.

The implications from the study of teacher-made tests are similar in nature. The findings indicate that a pupil's performance on new-type tests, in spite of the fact that the tests are objective in respect to scoring, is relatively variable. It is of significant practical value for persons using the new-type test to know this fact. If scores from tests are thought to be free from the effect of personal judgment, decisions based on them may not be checked by other evidence (as would tend to be the case when the decisions rest frankly upon personal opinion).

The facts here revealed, also, have certain implications for the theory or science of education. Since the appearance of Thorndike's *An Introduction to the Theory of Mental and Social Measurements* in 1904 educationists have tended to maintain that exact or absolute measurements are a fundamental requisite to a science of education. It is of value to know the relative extent to which so-called objective tests satisfy this requisite. The assumption that a science is possible only when uniform, relatively unvarying measures are available may or may not be sound. If it is sound, then the facts here reported indicate that education can hardly base its claim to be a science upon the reliability of the commonly used objective test.

Are new-type or objective tests more or less objective measuring instruments than are essay tests? The consensus of opinion on this point has been (if one may judge from the literature) that the new-type test is much more nearly free from the effects of personal factors. This opinion seems to have resulted from the fact that many writers have neglected to consider the complicated nature of the testing situation as a whole. Such evidence as is available indicates that there is probably little difference between the two types of test in respect to the presence of elements which cause disparity in the test results. However, this point was incidental to the present study; an adequate solution of this problem must await the accumulation of further evidence.

Finally, the educator may ask, In the light of the present evi-

dence, should new-type tests be used in the measurement of educational progress? An answer to this query can be made only when the purpose of the tester is known. Problems such as (a) the type of function measured by the new-type test, (b) the extent to which the use of such tests promotes reflective thought or rote learning, (c) the degree to which new-type tests are an index to various types of mental content, and other similar problems demand for their solution data in addition to those presented in this investigation.

PROBLEMS FOR FURTHER RESEARCH

As was stated in Chapter I (pp. 13-15), the purpose of this study was to discover factual evidence bearing upon definitely limited aspects of the large general problem of educational measurement. Therefore, numerous important and closely related problems of measurement are outside the scope of this investigation. Their solution requires specific researches designed to discover pertinent evidence. In order to facilitate such research, three major problems closely related to the present study are presented and discussed briefly.

1. *Causes of disparity.* What are the sources or causes of the disparity found to exist in the construction and use of the new-type or objective test? The subjectivity in the testing situation when new-type tests are used probably results from one or more of the following twelve causes,[2] each of which involves the personal judgment of the tester.

a. The items chosen. Teachers (or other test constructors) may and do choose different parts of given subject matter as suitable for test items. This difference in choice may be due to chance, to difference in judgment as to the importance of particular subject matter, to ease with which test items may be made, and perhaps to other factors. Further, it is doubtful whether the same teacher would select the same material for test items on two different occasions.

This problem of selecting test items involves the intricate theory of sampling. One may contend that the particular items chosen to make up a test do not significantly affect the nature of the test, provided the items are an adequate sample of the function tested. This contention seems to be based upon the assumption that one mental reaction is as likely to appear as another (quite as if one were sampling the apples in a barrel), and hence tends to ignore the possibility of mental organization. In the case of a particular individual, a spe-

[2] The probable sources of disparity listed here are regarded as more or less testable hypotheses.

cific item may recall a group of emotionally toned experiences which tend to block cognitive functioning and thus affect his performance on all other items in the test. In such a case, the fact that this particular item was chosen instead of numerous others that might have been chosen would have a significant effect upon the pupil's achievement rating.

b. Manner in which test items are stated. This factor involves the language in which the item is couched, the length of the item, and related aspects. For example, two true-false items may be constructed to test pupil acquaintance with a given bit of information, but a slight difference in language might make the difference between a correct and an incorrect response in the case of an individual child. In such a case, the personal decision of the tester to use a particular phrasing would be the cause of a change in the score of the pupil involved.

The problem may be more technically stated in the following manner. A test item is in essence (a) a series of symbols (b) organized into a unit (called a sentence) for the purpose of representing an idea. Two external factors determine the idea which a given unit conveys: First, the particular symbols used, and second, the organization of the symbols. Both of these factors are subject to wide variation from tester to tester. It follows, therefore, that when a given tester selects a specific set of symbols and organizes them in a specific manner in order to produce a test item, in both choice and organization of symbols he has used his personal judgment to a large degree. It follows, also, that the reaction of the testee to that item is conditioned in some degree by these subjective choices on the part of the tester.

c. Type of test item. Ruch[3] lists seventeen types of objective test item. There are several varieties of many of these types. When a test is constructed the decision must be made as to what type, or types, of item are to be used. Two test items intended to measure exactly the same information may produce different responses if the items are different in type. It seems highly probable that a multiple-choice test item measures a different function from that measured by a completion item. For example, suppose two teachers independently decide that a child in fifth grade geography should know what ocean a little Spanish boy would cross if he sailed directly to New York from his home in Spain. In order to determine whether or not the pupils have this information, each teacher constructs an

[3] Ruch, *op. cit.*, p. 189.

objective or new-type test item. The first teacher puts her item in the following form: A little boy coming directly to New York from his home in Spain would cross the (Pacific, Atlantic, Indian, Arctic) Ocean. (Correct response to be underlined.) The second teacher with exactly the same purpose in mind states her item in the following manner: A little boy coming directly to New York from his home in Spain would cross the ..... Ocean. (Correct response to be written in the blank space.) Are the two items really alike? Or are they different? Certainly the two items may (perhaps do) occasion different mental processes.[4] There is reason to believe that the difference in mental processes is sufficient to cause in many cases variation in pupil response.

d. Variety and proportion of types of item. A test may consist entirely of one type of item, or it may be composed of a number of types. There may be few or many items of any given type. A test made up of twenty multiple-choice and twenty true-false items may be quite different from a test which consists of five each of true-false, multiple-choice, completion, and matching items. The transition from one type of mental activity to another may affect the performance on particular items or on the test as a whole. In constructing a test the tester must inevitably exercise judgment as to the variety and proportion of items he will use. Differences in judgment at this point make for variation in responses.

e. Grouping and other relational aspects of items. In a given test the items which are based upon closely related aspects of knowledge may be grouped together, or the items may appear in the test without regard to their content. Whether easy or difficult items are placed toward the beginning or later in the test is a simple illustration of the problem here in question. If in the case of a given child the first five items on a test are very difficult, affective disturbance may influence the pupil's performance on the remaining part of the test. On the other hand, the pupil who answers the first five items with ease and confidence may approach the remainder of the test with an attitude such as will promote effective work. Thus, the mere decision to place particular items in a given position may have an effect upon pupil performance on the test.

The organization of the test items may, also, be of much importance. A test which requires a series of relatively isolated re-

[4] Weidemann has called attention to this point. Charles C. Weidemann, *How to Construct the True-False Examination* (Contributions to Education, No. 225; New York: Teachers College, Columbia University, 1926).

sponses may yield different results than a test composed of the same items organized into meaningful groups and arranged in systematic sequence. For example, ten consecutive items testing the pupil's knowledge of the industrial products of Germany may produce different results than the same ten items would if scattered throughout fifty other items. In the first case, the pupil's mental activity may tend to be organized or it may tend to be confused because of the proximity of many items on the same subject. Be that as it may, the results from a test may depend in some degree upon the tester's decision as to the organization of the items in his test.

f. Clarity and fullness of general and specific directions. Full explanation such as would tend to guarantee an understanding of the reaction desired may be given, or the directions may be brief to the extent of vagueness. Illustrations may or may not be given. The language in which directions are cast may vary in complexity and in clarity of expression. The pupils may, or may not, be permitted to ask questions before the test or during the test. It is possible that a single question asked by a student and answered by the teacher will affect the pupil's performance on an item or on the whole test.

g. Personality of the tester. The variation possible in the personality of the persons who administer a test is almost unlimited. For example, the tester may be very aloof and strict in manner, or he may be friendly and relatively lax in discipline. The first condition may produce a negativistic feeling on the part of some pupils, thus causing a poor quality of work; the same condition may cause other more phlegmatic pupils to do a type of work better than otherwise would have been the case. On the other hand, the second condition may promote or hinder efficiency in particular cases. Further, the same tester may manifest varying attitudes on different occasions, depending upon his general state of health, his experience immediately preceding, and the like.

h. Time allowed for taking test. Twenty items to be answered in twenty minutes may constitute quite a different test from a test made up of the same items to be answered in forty minutes. The general social atmosphere in which the test is taken may be rushed and tense, or the condition may be such as to promote the feeling of freedom and ease. Also, a given pupil may be able to respond effectively if he is not under tension and is given time to think, whereas, under opposite conditions, his performance may be poor. In the case of another pupil, the reverse may be true; that is, he may do his best work under the tension of strict requirements. Thus, the sub-

jective decision as to the time limits of a test may affect significantly the resulting pupil scores.

i. Number of items. The variation in respect to the length of the test is theoretically unlimited. A test composed of ten items constructed to cover ten pages of subject matter may not be comparable to a test made up of 150 similar items constructed for the same purpose. Once the teacher has decided to give a test, she must decide how many items are necessary for an adequate test. There is little reason to believe that two teachers will in a given case agree as to the number of items necessary for a good test.

j. Type of mental activity required by items. A test may be composed of items which require specific knowledge reactions, or the test may be composed principally of items which demand reflective thought. Further, a test may be made up of a combination of these two types of mental activity, as well as of many others. A test composed of twenty items, each of which requires the pupil to weigh a body of evidence and come to a conclusion, is very different from a test of twenty specific factual items based on the same subject matter. A given pupil, thus, may be able to do excellent work on the first test and very poor work on the second, or the reverse may be true. However, both tests may be considered as adequate measures of achievement in a given subject by the respective author of each test.

k. The evaluation of pupil responses. There is the problem of determining when the response of a child represents the mental content contemplated in the test item. The following item appeared in a history test: "Instead of states as in the United States the political divisions in Canada are *(provinces)*." The following are three responses made by as many children: (1) provinces, (2) provens, (3) profens. Is the mental content the same in each case? Does response (1) represent a better quality of mental reaction than does (2) or (3)?

One should note that the problem raised here is not restricted to completion exercises. On the contrary, the problem of truly evaluating a pupil response arises for every type of item. For example, when a pupil responds to a true-false item, only the product of his reaction is manifested; the mental activity which led to the response is completely hidden.

l. Pupil interpretation of items. The same item may be variously interpreted by different children or by the same child at different times. The child's previous experience, his physical and mental

state, his conative state and tone, and many other significant factors may affect his interpretation of a given item.

Since a pupil's performance on a given test varies with his interpretation of the items which make up the test, and since this interpretation is dependent upon a large number of intricate psychological factors, it follows that variation in pupil interpretation of test items may be an important cause of variation in test results.

2. *Possibility of absolute measurements.* Is it possible to develop measures in the social sciences which are both adequate and free from subjective elements? There are several positions which may be taken with relation to this question.

First, one may content that the testing situation is so complex as to frustrate the attempt to secure measures that take into account a sufficient number of the significant variables involved. One of the most difficult problems connected with psychological measurement grows out of the facts (a) that the reaction of that which is measured is basic to any measurement whatsoever, and (b) that the nature of this reaction depends upon a number of intricate factors. That is to say, in psychological measurement the instrument is an instrument of measurement only when it is reacted to by the thing being measured. The accurateness of the measurement is largely dependent upon the activity of the measured object. And further, the type and quality of activity of a given mind at a given moment depends upon a sensitive mental organization built up during the entire history of the mind concerned. This marked variability of the measured object causes variation in the results of measurement, although, objectively speaking, the instrument and the method of application remain constant.

An illustration may clarify this point. If one measures the height of a child, except for the fact that the child's body occupies space and is linear in nature, the accuracy of the measurement depends largely upon the nature of the instrument and the care with which it is applied. If the instrument (a steel tape) has been constructed according to certain generally accepted standards and if the tape is applied so as to avoid error, the reaction of the object measured is a relatively unimportant aspect of the measurement situation. However, the situation differs if one attempts to measure a child's knowledge of a process in arithmetic. Suppose ten examples are used as the instrument of measurement. In this case the reaction of the child is the basic aspect of the whole measurement situation. The child's performance is considered as an index to the

ability which presumably the performance represents. If for any one of numerous possible reasons the child does not respond "normally" significant error in the measurement results.

Further, a combination of specific reactions to specific situations may not be a meaningful index to the achievement of the individual as a whole. If (a) exact, quantitative measures are possible only when a measurable unitary phenomenon has been abstracted and isolated, (b) a person is essentially a whole rather than a sum of the "parts" of which he is constituted, and (c) the purpose of measurement is to ascertain the efficiency with which the whole organism adjusts itself to its environment, it follows then that when one measures the parts (as is so accurately done in the physical sciences) one may have done little or nothing toward securing a meaningful measure of the probable performance of the person as a whole.

Second, one may share the faith expressed by Thorndike in the following statement:

> We have faith that whatever people now measure crudely by mere descriptive words, helped out by the comparative and superlative forms, can be measured more precisely and conveniently if ingenuity and labor are set at the task. We have faith also that the objective products produced, rather than the inner condition of the person whence they spring, are the proper point of attack for the measurer, at least in our day and generation.
>
> This is obviously the same general creed as that of the physicist or chemist or physiologist engaged in quantitative thinking—the same, indeed, as that of modern science in general. And, in general, the nature of educational measurements is the same as that of all scientific measurements.[5]

Or finally one may take the closely related position that measurements of psychological entities comparable to measurements in the physical sciences are possible, but that such instruments must be developed very gradually as insights into the relationships requisite to the production of a refined instrument of measurement are gained. An extended quotation from the German psychologist and physicist, Wolfgang Köhler, may illuminate this point:

> The problems which Galileo attacked in the seventeenth century could be solved quantitatively at once, the qualitative experience of everyday life having sufficiently provided the necessary basis. But for the majority of psychological problems this is not the case. Where do we have that first more or less qualitative knowledge of important functional relationships in psychology which might become the basis for indirect and

[5] Thorndike, *Seventeenth Yearbook for the National Society for the Study of Education,* pp. 16-17.

exact measurement? It does not exist. Since the development of more "exact" methods presupposes its existence, our main task must be the gathering of that knowledge. In most cases our preliminary advance in this direction will have to be crude and qualitative. Whoever protests that conclusion in the name of purism does not understand our actual situation in psychology; he sees neither the nature of, nor the historical background prerequisite for, special quantitative methods. If we wish to imitate the physical sciences, we must not imitate them in their contemporary, most developed form; we must imitate them in their historical youth, when their state of development was comparable to our own at the present time. Otherwise we should behave like boys who try to copy the imposing manners of full-grown men without understanding their *raison d'être,* also without seeing that in development one cannot jump over intermediate and preliminary phases. A survey of the history of physics is certainly illuminating. Let us imitate the natural sciences, but intelligently!

Behavior is enormously rich in forms and nuances. Only acknowledging this wealth, and studying it directly as it is given in all its fascinating varieties, shall we become able gradually to find those forms of more quantitative, and perhaps more accurate, procedure which may become as adequate for us as are the methods of physics in its realm. At present, and in this broader historical perspective, qualitative observation and analysis may be, in a sense, more exact, i. e., adequate to our subject-matter, than much blind measurement. We shall press forward towards more refined methods, of course; but owing to our situation as beginners, we can go forward only through the use of less refined methods for the time being.[6]

3. *Desirability of absolute measurements.* If the assumption that absolute educational measurements are possible is granted for the moment, another problem[7] arises: Should educational measurements be objective in the sense in which the term is used in the physical sciences?[8] On the one hand, it is possible that intelligent personal judgment, based upon specific and locally determined educational aims, is in actual school procedures the most valuable type of measurement. If the statement just made is sound, the performance required and the meaning attributed to a performance would vary from community to community and probably from child to child. That is, achievement would be considered as a relative concept, its inter-

[6] Köhler, *op. cit.,* pp. 43-44. See also J. P. Brown, "A Methodological Consideration of the Problem of Psychometrics," *Erkenntnis,* iv Band (1934) Heft 1, pp. 46-61.

[7] The problem relating to the desirability of absolute measurements seems to be essentially a problem for research in the philosophy of education. The results of such studies should contribute much to the development of effective educational measurements.

[8] See quotations from Thorndike given on p. 21 for elucidation of the meaning of this type of measurement.

pretation in any given case depending upon the aims of the educator with respect to the factors of the specific situation. In essence this position rests upon the belief that the first and most basic problem of educational measurement is the ascertaining of educational aims; and that since tests are constructed to determine the extent to which the aims or objectives have been realized, the nature of the tests must grow out of the nature of the objectives.[9] Testing, according to this view, is essentially a process of gathering evidence upon which a decision as to the presence of the learning product may be based.[10]

On the other hand, progress in educational procedures may depend chiefly upon the development of exact and uniform measuring instruments which are capable of yielding accurate and unambiguous measures of educational achievement.[11] Perhaps the goal of test construction should be the production of universally applicable tests and scales in all subjects to the end that a unit of achievement in arithmetic at one developmental level (or for a given child), for example, may be comparable to a unit of achievement in that subject at any other level (or for any other child).

Finally, the positions stated in the two preceding paragraphs may be subject to effective synthesis. Eventually, uniformity may be attained in some educational objectives in such degree as to permit uniform standards. In these cases exact (in a limited sense) and nonpersonal measuring procedures may have a contribution to make to those aspects of educational measurements for which the most accurate (in terms of specific aims) measure involves a large portion of the measurer's personal judgment. For although this personal judgment must function within a frame of reference which is personal, it may be desirable that the judgments made *within this frame of reference* be as nonpersonal or objective as possible.

[9] The significance of objectives for adequate measurement has been very ably outlined and defended by R. W. Tyler. For extended and illuminating discussion see articles by Tyler appearing in *Educational Research Bulletin* (Columbus: Bureau of Educational Research, Ohio State University), Vols. XI to XIV, inclusive.

[10] H. C. Morrison, *The Practice of Teaching in the Secondary Schools* (Chicago: The University of Chicago Press, 1926), chap. v.

[11] For a critical discussion of this point see W. A. Brownell, "The Use of Objective Measures in Evaluating Instruction," *Educational Method*, XIII (May-June, 1934), 401-408.

APPENDIX

TABLE II. NUMBER AND TYPE OF ITEMS IN EACH OF SIXTY-THREE TEACHER-MADE OBJECTIVE TESTS

Test	Testing Unit	Number of Various Types of Item						Total Number of Items
		Completion	Matching	Multiple-Choice	True–False	Location	Unclassified	
IG5*.....	Identical	20	..	..	..	..	..	20
IIG5.....	"	10	..	..	10	..	..	20
IIIG5....	"	..	..	..	20	..	..	20
IVG5....	"	..	..	..	20	..	..	20
VG5.....	"	20	..	..	..	..	..	20
VIIG5....	"	20	..	..	40	..	..	60
VIIIG5...	"	10	..	..	10	..	..	20
IXG5....	"	15	..	..	20	..	..	35
XIIG5...	"	20	..	..	..	..	..	20
XIIIG5 ..	"	..	..	10	10	..	..	20
IG6.....	"	40	..	..	..	..	..	40
IIG6.....	"	15	..	..	25	..	..	40
IVG6.....	"	..	..	..	25	..	..	25
VIG6.....	"	9	..	5	10	..	..	24
VIIIG6...	"	5	..	..	15	..	..	20
IXG6....	"	..	..	33	..	..	..	33
XG6.....	"	5	..	5	10	..	..	20
XIG6....	"	10	..	..	15	..	..	25
XIIIG6 ..	"	7	..	..	8	..	..	15
IG7.....	Semester	5	15	5	30	18	16	89
IIG7.....	"	50	..	20	22	..	..	92
IIIG7....	"	..	..	25	8	..	..	33
IVG7.....	"	25	..	..	25	..	..	50
VG7.....	"	10	..	15	20	..	..	45
IH5.....	Identical	..	..	16	8	..	..	24
IIIH5....	"	..	..	..	25	..	..	25
IVH5....	"	5	10	..	5	..	..	20
VH5.....	"	20	..	..	..	..	..	20
VIIH5....	"	10	..	10	..	..	..	20
VIIIH5...	"	12	..	8	10	..	..	30
IXH5....	"	..	..	..	25	..	..	25
XIIH5...	"	25	..	..	..	..	..	25
IH6.....	"	10	..	20	10	..	..	40
IIH6.....	"	25	..	..	..	..	..	25
IVH6....	"	10	..	..	10	..	..	20
IVH6A...	"	20	..	..	..	..	..	20
VIH6....	"	9	5	..	10	..	..	24
VIIIH6...	"	11	..	11	..	..	..	22
XH6.....	"	17	..	..	..	..	..	17
XIH6....	"	15	..	..	10	..	..	25
XIIIH6..	"	..	..	..	65	..	..	65
IH7a.....	Semester	5	..	5	15	..	..	25
IH7b.....	"	5	..	5	15	..	..	25
IH7c.....	"	5	..	5	15	..	..	25
IIH7a....	"	10	25	..	15	..	..	50
IIH7b....	"	10	30	..	10	..	..	50
IIH7c....	"	10	30	..	10	..	..	50

*I, Teacher I; G, Geography; 5, Grade 5.

TABLE II. (Continued)

Test	Testing Unit	Number of Various Types of Item						Total Number of Items
		Completion	Matching	Multiple-Choice	True–False	Location	Unclassified	
IIH7d....	"	10	30	..	10	..	..	50
IIH7e....	"	10	30	..	10	..	..	50
IIIH7....	"	25	..	..	25	..	..	50
IVH7....	"	10	30	20	40	..	..	100
VH7.....	"	20	..	20	30	..	..	70
VIH7....	"	10	20	10	10	..	..	50
IHSF....	"	25	..	25	50	..	..	100
IIHSF ...	"	30	20	20	10	..	..	80
IIHSB....	"	20	15	30	40	..	..	105
IIHSB...	"	20	20	..	25	..	..	65
IHSH....	Identical	23	15	16	25	..	..	69
IIHSH...	"	15	18	..	20	..	..	53
IIIHSH..	"	12	15	12	20	..	..	59
IVHSH...	"	8	15	..	16	..	..	39
IHe6.....	"	10	..	5	10	..	..	25
IIHe6....	"	..	..	..	25	..	..	25

TABLE XLV. ARITHMETIC REASONING—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	Test *a* From Test *b*		Test *a* From Test *c*		Test *b* From Test *c*	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
36–38.....	..	....	..	....	1	0.22
33–35.....	..	....	..	....	1	0.44
30–32.....	..	....	1	0.22	2	0.87
27–29.....	1	0.22	5	1.31	1	1.09
24–26.....	2	0.65	4	2.18	3	1.74
21–23.....	1	0.87	8	3.92	8	3.48
18–20.....	8	2.61	14	6.96	18	7.39
15–17.....	12	5.22	26	12.61	24	12.61
12–14.....	26	10.87	37	20.65	48	23.05
9–11.....	52	22.17	61	33.91	77	39.79
6–8	89	41.52	96	54.78	94	60.22
3–5	149	73.91	117	80.22	85	78.70
0–2	120	100.00	91	100.00	98	100.00
Total.....	460		460		460	
Mean.....	6.3		8.0		8.4	

TABLE XLVI. GEOGRAPHY—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST *a* FROM TEST *b*		TEST *a* FROM TEST *c*		TEST *b* FROM TEST *c*	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
42–44	..	....	1	0.22	..	....
39–41	..	....	4	1.09	2	0.43
36–38	..	....	6	2.40	..	0.43
33–35	1	0.22	11	4.79	..	0.43
30–32	5	1.31	17	8.49	8	2.17
27–29	5	2.40	21	13.06	15	5.43
24–26	10	4.57	26	18.71	8	7.17
21–23	15	7.83	21	23.28	18	11.09
18–20	18	11.75	33	30.45	20	15.44
15–17	29	18.05	41	39.36	30	21.96
12–14	37	26.09	56	51.53	41	30.87
9–11	68	40.87	52	62.83	60	43.92
6–8	74	56.96	64	76.74	84	62.18
3–5	109	80.65	50	87.61	91	81.96
0–2	89	100.00	57	100.00	83	100.00
Total	460		460		460	
Mean	9.0		14.1		10.0	

TABLE XLVII. HEALTH—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *b* AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST *b* FROM TEST *c*	
	Number of Pupils	Cumulative Percentage Distribution
48 — 50	1	0.22
45 — 47	..	0.22
42 — 44	..	0.22
39 — 41	3	0.87
36 — 38	3	1.52
33 — 35	3	2.17
30 — 32	6	3.48
27 — 29	9	5.44
24 — 26	12	8.05
21 — 23	32	15.00
18 — 20	40	23.69
15 — 17	41	32.60
12 — 14	54	44.34
9 — 11	60	57.39
6 — 8	68	72.17
3 — 5	59	85.00
0 — 2	69	100.00
Total	460	
Mean	12.1	

TABLE XLVIII. HISTORY—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST *a* FROM TEST *b*		TEST *a* FROM TEST *c*		TEST *b* FROM TEST *c*	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
42–44	..	....	2	0.43	..	....
39–41	..	....	..	0.43	..	....
36–38	..	....	1	0.65	..	....
33–35	..	....	5	1.74	4	0.87
30–32	1	0.22	2	2.17	4	1.74
27–29	..	0.22	13	5.00	10	3.91
24–26	4	1.09	21	9.57	12	6.52
21–23	6	2.39	34	16.96	22	11.30
18–20	21	6.97	31	23.69	30	17.82
15–17	36	14.80	65	37.83	51	28.91
12–14	55	26.76	66	52.17	56	41.08
9–11	84	45.02	73	68.04	59	53.91
6–8	87	63.92	51	79.13	81	71.52
3–5	88	83.05	53	90.65	75	87.83
0–2	78	100.00	43	100.00	56	100.00
Total	460		460		460	
Mean	8.8		13.2		11.3	

TABLE XLIX. LANGUAGE USAGE—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST *a* FROM TEST *b*		TEST *a* FROM TEST *c*		TEST *b* FROM TEST *c*	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
60–62	..	....	..	....	1	0.22
....	..	....	..	....	..	....
45–47	..	....	..	....	2	0.65
42–44	..	....	..	....	2	1.08
39–41	..	....	..	....	2	1.51
36–38	..	....	..	....	5	2.60
33–35	2	0.43	2	0.43	9	4.56
30–32	15	3.69	7	1.95	11	6.95
27–29	16	7.17	12	4.56	22	11.73
24–26	12	9.78	20	8.91	13	14.56
21–23	26	15.43	36	16.74	42	23.69
18–20	45	25.21	38	25.00	39	32.17
15–17	57	37.60	52	36.30	36	40.00
12–14	49	48.25	57	48.69	42	49.13
9–11	58	60.86	65	62.83	39	57.61
6–8	74	76.95	54	74.57	52	68.91
3–5	55	88.91	56	86.74	90	88.48
0–2	51	100.00	61	100.00	53	100.00
Total	460		460		460	
Mean	12.7		12.4		13.7	

TABLE L. LITERATURE—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a* AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST *a* FROM TEST *c*	
	Number of Pupils	Cumulative Percentage Distribution
27 — 29	5	1.09
24 — 26	16	4.57
21 — 23	22	9.35
18 — 20	38	17.61
15 — 17	41	26.52
12 — 14	53	38.04
9 — 11	69	53.04
6 — 8	76	69.56
3 — 5	80	86.95
0 — 2	60	100.00
Total	460	
Mean	10.7	

TABLE LI. READING—PARAGRAPH MEANING—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	TEST *a* FROM TEST *b*		TEST *a* FROM TEST *c*		TEST *b* FROM TEST *c*	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
45–47	1	0.22	..	....	..	....
42–44	1	0.44	..	....	..	....
39–41	1	0.66	..	....	..	....
36–38	3	1.31	..	....	..	....
33–35	1	1.53	5	1.09	3	0.65
30–32	3	2.18	1	1.31	1	0.87
27–29	6	3.49	5	2.40	2	1.30
24–26	16	6.97	13	5.23	10	3.47
21–23	10	9.14	17	8.92	5	4.56
18–20	36	16.97	38	17.18	19	8.69
15–17	45	26.75	49	27.83	38	16.95
12–14	46	36.75	42	36.96	59	29.78
9–11	71	52.18	50	47.83	60	42.83
6–8	68	66.96	77	64.56	62	56.31
3–5	92	86.96	88	83.69	120	82.39
0–2	60	100.00	75	100.00	81	100.00
Total	460		460		460	
Mean	10.8		10.4		8.9	

TABLE LII. SPELLING—GRADE-EQUIVALENT DISPARITY BETWEEN TESTS *a*, *b*, AND *c* IN TERMS OF MONTHS

Months of Disparity	Test *a* From Test *b*		Test *a* From Test *c*		Test *b* From Test *c*	
	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution	Number of Pupils	Cumulative Percentage Distribution
39–41	..	....	..	....	1	0.22
36–38	..	....	..	....	..	0.22
33–35	..	....	..	....	..	0.22
30–32	..	....	..	....	1	0.44
27–29	2	0.43	..	....	..	0.44
24–26	..	0.43	1	0.22	1	0.66
21–23	4	1.30	2	0.65	3	1.31
18–20	5	2.39	4	1.52	3	1.96
15–17	9	4.35	8	3.26	10	4.13
12–14	33	11.52	16	6.74	17	7.83
9–11	61	24.78	22	11.52	44	17.39
6–8	99	46.30	70	26.74	114	42.17
3–5	126	73.69	157	60.87	123	68.91
0–2	121	100.00	180	100.00	143	100.00
Total	460		460		460	
Mean	6.5		4.8		5.9	

TABLE LIII. WORD MEANING—DISPARITY BETWEEN TESTS *a* AND *c* IN TERMS OF MONTHS

Months of Disparity	Test *a* From Test *c*	
	Number of Pupils	Cumulative Percentage Distribution
33 — 35	1	0.22
30 — 32	..	0.22
27 — 29	1	0.44
24 — 26	1	0.66
21 — 23	3	1.31
18 — 20	20	5.65
15 — 17	28	11.73
12 — 14	32	18.68
9 — 11	51	29.76
6 — 8	75	46.10
3 — 5	131	74.57
0 — 2	117	100.00
Total	460	
Mean	7.2	

BIBLIOGRAPHY

NOTE: This bibliography consists only of titles that have been referred to in the text. Several extensive bibliographies, both on the general subject of educational measurement and on the more limited subject of objective or new-type tests, are available. These more complete bibliographies appear as starred (*) references in the list which follows.

*"Bibliography on Educational Tests and Their Use," *Review of Educational Research,* III (Feb., 1933), 62-80.

Brown, J. F. "A Methodological Consideration of the Problems of Psychometrics," *Erkenntnis,* iv Band (1934) Heft 1, pp. 46-61.

Brownell, William A. "On the Accuracy with Which Reliability May Be Measured by Calculating Test Halves," *Journal of Experimental Education,* I (March, 1933), 204-215.

———"The Use of Objective Tests in Evaluating Instruction," *Educational Method,* XIII (May-June, 1934), 401-408.

Chadwick, E. B. "Statistics of Educational Results," *The Museum, A Quarterly Magazine of Educational Literature and Science,* III (Jan., 1864), 479-484.

Dingle, Herbert. *Science and Human Experience.* New York: The Macmillan Company, 1932. 141 pp.

Foran, T. G., and Loyes, Sister M. Edmund. "The Relative Difficulty of Three Achievement Examinations," *The Journal of Educational Psychology,* XXVI (March, 1935), 218-222.

Garrett, Henry E. *Statistics in Psychology and Education.* New York: Longmans, Green, and Company, 1926. 317 pp.

Johnson, F. W. "A Study of High School Grades," *School Review,* XIX (Jan., 1911), 13-24.

*Jones, Vernon, and Brown, Robert H. "Educational Tests," *Psychological Bulletin,* XXXII (July, 1934), 473-499.

Kelly, F. J. *Teachers' Marks, Their Variability and Standardization.* Teachers College Contributions to Education, No. 66. New York: Teachers College, Columbia University, 1914. 139 pp.

Köhler, Wolfgang. *Gestalt Psychology.* New York: Horace Liveright, Inc., 1929. 403 pp.

*Lee, J. Murray, and Symonds, P. M. "New-Type or Objective Tests: A Summary of Recent Investigations," *The Journal of Educational Psychology,* XXIV (Jan., 1933), 21-38.

*———"New-Type or Objective Tests: A Summary of Recent Investigations," *The Journal of Educational Psychology,* XXV (March, 1934), 185-191.

McCall, William A. *How to Measure in Education.* New York: The Macmillan Company, 1922. 416 pp.

Monroe, Walter Scott. *An Introduction to the Theory of Educational Measurements.* New York: Houghton Mifflin Company, 1923. 364 pp.

Monroe, W. S., and Souders, L. B. *Present Status of Written Examinations and Suggestions for Their Improvement.* Bulletin No. 17. Ur-

bana: College of Education, Bureau of Educational Research, University of Illinois, 1923. 77 pp.

Morrison, Henry C. *The Practice of Teaching in the Secondary School.* Chicago: The University of Chicago Press, 1926. 661 pp.

Odell, C. W. *Educational Measurement in High School.* New York: The Century Company, 1930. 641 pp.

———— *A Glossary of Three Hundred Terms Used in Educational Measurement and Research.* Bulletin No. 40. Urbana: College of Education, Bureau of Educational Research, University of Illinois, 1928. 68 pp.

*———— *A Selected Annotated Bibliography Dealing with Examinations and School Marks.* Bulletin No. 43. Urbana: College of Education, Bureau of Educational Research, University of Illinois, 1929. 42 pp.

Pullias, E. V. "A Study of Current Opinion Concerning Objective Tests," *Educational Method,* XVI (April, 1937), 348-356.

Rice, J. M. "The Futility of the Spelling Grind," *Forum,* XXIII (June, 1897), 409-419.

———— "The Futility of the Spelling Grind," *Forum,* XXIII (April, 1897), 163-172.

Ritchie, A. D. *Scientific Method: An Inquiry into the Character and Validity of Natural Laws.* London: Kegan Paul, 1923. 204 pp.

Robinson, Daniel Sommer. *The Principles of Reasoning: An Introduction to Logic and Scientific Method.* New York: D. Appleton and Company, 1924. 393 pp.

*Ruch, G. M. *The Objective or New-Type Examination.* New York: Scott, Foresman and Company, 1929. 478 pp.

Ruch, G. M., De Graff, M. H., Gordon, W. E. . . . (and others). *Objective Examination Methods in the Social Studies.* New York: Scott, Foresman and Company, 1926. 116 pp.

Russell, Charles. *Standard Tests.* Boston: Ginn and Company, 1930. 516 pp.

Sandon, Frank. "The Necessary Imperfections of an Examination," *The British Journal of Educational Psychology,* V (June, 1935), 180-192.

*Smith, Henry Lester, and Wright, Wendell William. *Second Revision of the Bibliography of Educational Measurements.* Bulletin of the School of Education, Vol. IV, No. 2. Bloomington: Bureau of Co-operative Research, Indiana University, 1927. 251 pp.

*———— *Tests and Measurements.* New York: Silver, Burdett and Company, 1928. 540 pp.

Starch, Daniel. *Educational Measurements.* New York: The Macmillan Company, 1918. 202 pp.

Starch, Daniel, and Elliott, E. C. "The Reliability of Grading High School Work in Mathematics," *School Review,* XXI (April, 1913), 254-259.

Thorndike, Edward L. *An Introduction to the Theory of Mental and Social Measurements.* (Revised Edition, 1913.) New York: Teachers College, Columbia University, 1904. 277 pp.

———— "The Nature, Purposes, and General Methods of Measurements of Educational Products," *The Measurement of Educational Products.* Seventeenth Yearbook of the National Society for the Study of Edu-

cation, Part II. Bloomington, Illinois: Public School Publishing Company, 1918. Pp. 16-24.

Tyler, R. W. "Formulating Objectives for Tests," *Educational Research Bulletin,* Columbus, Bureau of Educational Research, Ohio State University, XII (Oct. 11, 1933), 197-206.

Weidemann, Charles C. *How to Construct the True-False Examination.* Contributions to Education, No. 225. New York: Teachers College, Columbia University, 1926. 118 pp.

Willoughby, R. R. "The Concept of Reliability," *Psychological Review,* XLII (March, 1935), 153-165.

Date Due

378.756 D877R no.1-5 229853

Duke University Libraries
D01129207M

D01129207M
Duke Library Service Center